INSTITUTO TECNOLÓGICO DE CD. MADERO

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

"Clasificación de Redes Complejas Usando Funciones de Caracterización que Permitan Discriminar entre Redes Aleatorias, Power-Law y Exponenciales"

PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS
EN CIENCIAS DE LA COMPUTACIÓN

P R E S E N T A:

I.S.C. Tania Turrubiates López

DIRECTOR:

MC. Claudia Guadalupe Gómez Santillán

CODIRECTOR:

MC. Rogelio Ortega Izaguirre

CD. MADERO, TAM. MEXICO NOVIEMBRE DEL 2007

INSTITUTO TECNOLÓGICO DE CD. MADERO

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

"Clasificación de Redes Complejas Usando Funciones de Caracterización que Permitan Discriminar entre Redes Aleatorias, Power-Law y Exponenciales"

PARA OBTENER EL GRADO DE:
MAESTRO EN CIENCIAS
EN CIENCIAS DE LA COMPUTACIÓN

P R E S E N T A:
I.S.C. Tania Turrubiates López

DIRECTOR:
MC. Claudia Guadalupe Gómez Santillán

CODIRECTOR:
MC. Rogelio Ortega Izaguirre

JURADO:

Presidente:	**Dra. Laura Cruz Reyes**
Secretario:	**Dr. Santos Aguilar De Luna**
Vocal:	**M.C. Claudia Guadalupe Gómez Santillán**
Suplente:	**M.C. Rogelio Ortega Izaguirre**

CD. MADERO, TAM. MEXICO NOVIEMBRE DEL 2007

D.I.

Instituto Tecnológico de Cd. Madero

Cd. Madero, Tam., a 05 de Noviembre de 2007.

Área: Posgrado
Nº Oficio: U5.412/07
Asunto: Autorización de Impresión de Tesis

C. ING. TANIA TURRUBIATES LÓPEZ
P r e s e n t e.

Me es grato comunicarle que después de la revisión realizada por el Jurado designado para su examen de grado de Maestra en Ciencias en Ciencias de la Computación, se acordó autorizar la impresión de su tesis titulada:

"CLASIFICACIÓN DE REDES COMPLEJAS USANDO FUNCIONES DE CARACTERIZACIÓN QUE PERMITAN DISCRIMINAR ENTRE REDES ALEATORIAS, POWER – LAW Y EXPONENCIALES"

Es muy satisfactorio para la División de Estudios de Posgrado e Investigación compartir con Usted el logro de esta meta. Espero que continúe con éxito su desarrollo profesional y dedique su experiencia e inteligencia en beneficio de México.

A t e n t a m e n t e
"POR MI PATRIA Y POR MI BIEN"

Ma. Yolanda Chávez Cinco
M.P. María Yolanda Chávez Cinco
Jefa de la División

S.E.P.
DIVISION DE ESTUDIOS DE POSGRADO E INVESTIGACION
ITCM

MYCHC ' NLCO ' cerc*

DEDICATORIA

A mis padres:
Por enseñarme que la perseverancia y la confianza
son las principales herramientas para lograr mis objetivos.

A mis hermanos:
Por el apoyo y motivación que siempre me han brindado
y la paciencia que desde niña me han tenido.

A Moyito:
Por enseñarle a su tía que después de todo una distracción,
despeja la mente y así llegan mejores ideas.

Gracias Dios por iluminar mi camino
y tenerme siempre bajo tu cobijo.

AGRADECIMIENTOS

Mi profundo agradecimiento a los miembros del comité tutorial de esta tesis: Dra. Laura Cruz Reyes, Dr. Santos Aguilar de Luna, M.C. Claudia Guadalupe Gómez Santillán, M.C, Rogelio Ortega Izaguírre por las sugerencias dadas durante el desarrollo de esta tesis.

Mi sincero aprecio a M.C. Claudia Guadalupe Gómez Santillán y M.C. Rogelio Ortega Izaguirre por haber dirigido esta tesis; y la Dra. Elisa Schaeffer por las aportaciones realizadas a este trabajo.

Reciban mi reconocimiento las instituciones de las cuales recibí apoyo: el Instituto Tecnológico de Ciudad Madero (ITCM), el Consejo Nacional de Ciencia y Tecnología (CONACYT) quienes proporcionaron todas las facilidades necesarias para el desarrollo de esta investigación, a las autoridades del Instituto Tecnológico Superior de Álamo Temapache (ITSAT) por otorgar el permiso para realizar los estudios de maestría.

Reciban mi agradecimiento a los residentes y ayudantes de investigador que colaboraron en este proyecto.

Doy gracias a mis compañeros de maestría por el compañerismo y amistad que me brindaron y a mis compañeros y amigos del ITSAT que me brindaron su soporte moral y amistad. Para ellos mi estimación.

ARTÍCULOS REALIZADOS

Los resultados y aportaciones de esta investigación han sido presentados conjuntamente con mis asesores en los siguientes artículos internacionales:

- **"Impact of Dynamic Growing on the Internet Degree Distribution"**, ISPA'07, Fifth International Symposium on Parallel and Distributed Processing and Applications, Ontario, Canada, 29 - 31 de Agosto de 2007. Lecture Notes in Computer Science.
- **"Experimental Design for Selection of Characterization Functions that allow Discriminate among Random, Scale Free and Exponential Networks"**, ACS'07, 14th International Multi-conference Advanced Computer System, Międzyzdroje, Poland, 17 – 19 de Octubre de 2007. Polish Journal of Environmental Studies, Vol. 16, No. 5B, pp. 67 – 71.
- **"Studying the Impact of Growing Dynamic in the Internet Topology"** ACS'07, 14th International Multi-conference Advanced Computer System, Międzyzdroje, Polonia, 17 – 19 de Octubre de 2007. Polish Journal of Environmental Studies, Vol. 16, No. 5B, pp. 117 – 120.
- **"Statistical selection of relevant features to classify random, scale-free and exponential networks"**, HAIS'07, 2nd International Workshop on Hybrid Artificial Intelligence System, Salamanca, España, 12, 13 de Noviembre de 2007. Innovations in Hybrid Intelligent Systems, Series Advanced in Soft Computing Vol. 44.
- **"Generación y Validación de Redes Complejas con Diferencia de Grado"**, 1er. Congreso Estudiantil de Investigación Multidisciplinaria 2006, 30 de Mayo de 2006, Universidad Valle del Bravo Campus Tampico.

En foros nacionales se han presentado los siguientes artículos:

- **"Análisis de la Distribución del Grado en la Topología de Internet"**, CIICC'06, 13avo. Congreso Internacional de Investigación en Ciencias Computacionales, 15 - 17 de Noviembre del 2006, Instituto Tecnológico de Ciudad Madero.
- **"Statistical selection of relevant features to classify random, scale-free and exponential networks"**, MICAI'07, 6th Mexican International Conference on Artificial Intelligence, Aguascalientes, México, 4-10 de Noviembre de 2007. Research in Computing Science.
- **"Diseño Experimental para la Selección de Funciones de Caracterización que permiten discriminar entre Redes Aleatorias, de Libre Escala y Exponenciales"** CIICC'07, 14avo. Congreso Internacional de Investigación en Ciencias Computacionales, Orizaba, México, 7-9 de Noviembre de 2007.

RESUMEN

Los sistemas complejos pueden ser modelados mediante grafos, conocidos como redes complejas. Las redes complejas poseen una estructura topológica no trivial, lo que ha motivado el estudio de características topológicas de redes del mundo real. El conocimiento de estas características puede ser usado para optimizar el desempeño de los procesos que en ellas se llevan a cabo, tales como la búsqueda de recursos distribuidos, administración de tráfico y diseño de algoritmos de enrutamiento.

En este trabajo, el problema de la clasificación de redes complejas usando funciones de caracterización fue abordado. Esto es, dado un conjunto de redes complejas de diferente tipo y un conjunto de funciones de caracterización que permiten estudiar características topológicas de la red, la tarea es identificar cuáles son las funciones permiten de manera cuantitativa clasificar ese conjunto de redes. Hasta ahora, el método para identificar el tipo de red es observando la gráfica de la distribución del grado. Algunos investigadores se han enfocado en clasificar redes del mundo real mediante funciones de caracterización dentro de un tipo de red especifico, sin mostrar evidencias de un análisis detallado de las funciones que pudiera determinar sí el conjunto de funciones utilizadas son suficientes para lograr una clasificación eficiente o cuáles son las funciones que llevan a mejores resultados de clasificación.

En esta tesis, se desarrolló una metodología que toma como base la arquitectura de un agente de aprendizaje para identificar, por medio de un diseño experimental y una serie de pruebas estadísticas, el conjunto mínimo de funciones de caracterización relevantes y no redundantes que permitan discriminar cuantitativamente entre diferentes tipos de redes complejas como las redes Aleatorias, Power-Law y Exponenciales. Los resultados de esta investigación muestran que la función de caracterización *Coeficiente de Dispersión del Grado* (*DDC*) permite cuantitativamente discriminar entre redes Aleatorias, Power-Law y Exponenciales. La exactitud del proceso de clasificación usando esta función como entrada, es del 99.78% en el conjunto de instancias generadas para este trabajo y 91.25% con instancias de las cuales no se sabe su naturaleza y que fueron generadas independientemente por otros investigadores.

SUMMARY

The complex systems can be modeled by the means of graphs, known as complex networks. Complex networks have a non-trivial topological structure, which has motivated the study of topological characteristics of real-world networks. Knowledge on such characteristics can be used to optimize the performance of processes carried out in these networks, such as the search of distributed resources, traffic management and design of routing algorithms

In this work, the problem of the classification of complex networks using characterization functions is treated. That is, given a set of complex networks of different type and a set of characterization functions that allow the study of the topological characteristics of networks, the task is to identify which of the functions allow to classify in a quantitative way that set of networks. Until now, the method to identify the type of a complex network is by observing the plot of the degree distribution. Some researchers have focused in classifying complex networks of the real world through characterization functions within a specific class of complex networks, without showing evidence of detailed analysis of the functions that would allow either determining if the set of functions used is sufficient for carrying out an efficient discrimination or which of the functions perform better in obtaining the best results of classification.

In this work, a methodology based on the learning-agent architecture was developed to identify, by the means of an experimental design and a series of statistical tests, the minimal set of characterization functions that are relevant and non-redundant that allows to discriminate quantitatively among different types of complex networks, such as the random, power-law, and exponential networks.

The results of this research show that the *Degree Dispersion Coefficient* (*DDC*) allows to discriminate quantitatively among different types of complex networks, as the random, power-law, and exponential networks. The correctness of the classification process using this function as input is 99.78% on a set of pre-labelled instances generated for this work and 91.25% on a blind set of instances generated by another researcher independently.

TABLA DE CONTENIDO

Lista de Tablas iix
Lista de Figuras x
Lista de Ecuaciones xii
Lista de Cuadros xiii

Capítulo 1. INTRODUCCIÓN 1

1.1 Definición del Problema 2
1.2 Definición Formal 3
1.3 Justificación 3
1.4 Objetivos 4
1.5 Hipótesis 5
1.6 Organización del Documento 5

Capítulo 2. REDES COMPLEJAS 6

2.1 Teoría de Grafos 6
2.2 Redes Complejas 8
2.3 Funciones de Caracterización 9
2.3.1 Distribución del Grado 10
2.3.2 Coeficiente de Agrupamiento 11
2.3.3 Longitud de Ruta más Corta Característica 11
2.3.4 Diámetro de la Red 12
2.3.5 Eficiencia de la Red 12
2.3.6 Coeficiente de Dispersión del Grado 13
2.4 Tipos de Redes Complejas 14
2.4.1 Redes Aleatorias 14
2.4.2 Redes Power-Law 16
2.4.3 Exponenciales 17
2.4.4 Small World 18
2.5 Modelos de Generación de Redes 19
2.5.1 Modelos de Generación sin Crecimiento 19
2.5.1.1 Modelo de Erdös y Rényi (ER) 19
2.5.2 Modelos de Generación Basados en Crecimiento: 19
2.5.2.1 Modelo Barabási-Albert (BA) 20
2.5.2.2 Modelo de Liu 20

Capítulo 3. APRENDIZAJE AUTOMÁTICO 22

3.1 Agente de Aprendizaje 22
3.2 Aprendizaje Automático 23
3.2.1 Tipos de Aprendizaje 23
3.2.1.1 Aprendizaje Supervisado 24
3.2.1.1.1 Separabilidad de Clases 25
3.2.2 Técnicas de Aprendizaje 27
3.3 Selección de Características 29

Capítulo 4. ANÁLISIS Y DISEÑO DE EXPERIMENTOS 32

4.1 Análisis y Diseño De Experimentos 32
4.1.1 Diseño Factorial 34
4.2 Métodos Multivariados 37
4.2.1 Análisis de la Matriz de Correlación. 38
4.2.2 Análisis Multivariado de la Varianza. 39
4.2.3 Análisis Discriminante 40
4.3 Determinación del Tamaño de la Muestra. 40

Capítulo 5. ESTADO DEL ARTE 42

5.1 Trabajos Relacionados 42
5.2 Discusión de los Trabajos Relacionados 44
5.2.1 Análisis Comparativo 45

Capítulo 6. METODOLOGÍA 47

6.1 Metodología Propuesta 47
6.2 Generación de Redes Complejas 49
6.2.1 Generación de Instancias De Red. 50
6.2.2 Extracción de Características Topológicas. 51
6.3 Selección de Características 52
6.3.1 Identificación de Características Relevantes e Irrelevantes. 54
6.3.2 Identificación de Características Muy Relevantes y Relevantes Débiles 55
6.3.3 Eliminación de Características Redundantes. 55
6.3.4 Identificación del Conjunto Óptimo 55
6.4 Clasificación de Redes Complejas 56
6.4.1 Identificación del Clasificador con el Mejor Desempeño. 56
6.4.2 Clasificación de Instancias de Naturaleza Desconocida 56
6.4.3 Determinación del Tipo de Red en cada Nodo de las Instancias de Internet . 57

Capítulo 7. EXPERIMENTACIÓN Y ANÁLISIS DE RESULTADOS 58

7.1 Generación de Redes Complejas 59
7.1.1 Determinación del Número de Instancias de Red. 60
7.2 Selección de Características 63
7.2.1 Identificación de Características Relevantes e Irrelevantes. 66
7.2.1.1 Diseño Factorial Modelo de los Efectos Fijos 68
7.2.1.2 Diseño Factorial Modelo Mixto 70
7.2.1.3 Análisis de las Gráficas de los Efectos. 73
7.2.1.4 Análisis de las Gráficas de los Residuales 75
7.2.2 Identificación de Características Muy Relevantes y Relevantes Débiles 77
7.2.3 Eliminación de Características Redundantes. 78
7.2.4 Identificación del Conjunto Mínimo 79
7.3 Clasificación de Redes Complejas 82
7.3.1 Identificación del Clasificador con el Mejor Desempeño 83
7.3.2 Clasificación de Instancias de Naturaleza Desconocida 84
7.3.3 Determinación del Tipo de Red en cada Nodo de las Instancias de Internet . 85

Capítulo 8. CONCLUSIONES Y TRABAJOS FUTUROS 88
8.1 Conclusiones 88
8.2 Trabajos Futuros 90
Anexo A 92
A1. Instancias de Redes Reales 92
Anexo B 94
B1. Histogramas de las Funciones de Caracterización según el Tipo de Red 94
B2. Histogramas de las Funciones de Caracterización según el Número de Nodos 98
Anexo C 103
C1. Resultados del Diseño Factorial Modelo de los Efectos Fijos. 103
C2. Resultados del Diseño Factorial Modelo Mixto 106
Anexo D 110
D1. Gráficas de los Efectos Principales y de la Interacción de Factores 110
Anexo E 117
E1. Gráficas de los Residuales para cada Función de Caracterización 117
Anexo F 126
Anexo G 129
Anexo H 134
Referencias 138

LISTA DE TABLAS

4.1 Arreglo general de un diseño factorial de dos factores 36
5.1 Análisis comparativo de los trabajos relacionados 45
6.1 Arquitectura de un agente de aprendizaje vs. Clasificación de redes complejas 48
7.1 Fórmulas para determinar del tamaño de la muestra 60
7.2 Determinación del tamaño de la muestra para el modelo de efectos fijos 61
7.3 Determinación del tamaño de la muestra para el modelo mixto 63
7.4 Estadísticos descriptivos de la función *DDC*(*G*) por tipo de red. 65
7.5 Estadísticos descriptivos de la función *DDC*(*G*) por número de nodos. 66
7.6 Hipótesis asociadas al modelo de los efectos fijos 68
7.7 Valores F_0 calculados mediante GLM para cada función de caracterización. 69
7.8 Valores F_0 calculados por la prueba MANOVA y valores de la distribución F. 69

7.9 Hipótesis asociadas al modelo mixto 70
7.10 Valores F_0 calculados mediante GLM para cada función de caracterización. 71
7.11 Componentes de la varianza para cada función de caracterización. 72
7.12 Valores F_0 calculados por la prueba MANOVA y valores de la distribución *F*. 73
7.13 Resultados del Análisis Discriminante Cuadrático. 80
7.14 Pruebas estadisticas utilizadas en la selección de caracteristicas. 82
7.15 Información de las instancias de Interner para los años 1997, 2000 y 2003 85
g1. Resultado de la clasificación de redes complejas de naturaleza desconocida 129
h1. Métodos y algoritmos para la selección de características usados para comparar los resultados obtenidos con el enfoque estadístico 134
h2. Comparación de los resultados de la selección de características efectuada mediante métodos de análisis multivariado, algoritmos de aprendizaje y el enfoque estadístico propuesto 135

LISTA DE FIGURAS

1.1 Planteamiento del problema 2
2.1 Grafo y su correspondiente matriz de adyacencia 7
2.2 Grafos con la misma topología pero diferente forma 8
2.3 Extracción de características topológicas mediante funciones de caracterización 10
2.4 Red aleatoria: a) su estructura tipica, b) gráfica de la distribución del grado. 16
2.5 Red power-law: a) su estructura tipica, b) gráfica de la distribución del grado. 17
2.6 Red exponencial: a) su estructura tipica, b) gráfica de la distribución del grado. 18
3.1 Arquitectura de un agente con capacidad de aprender 23
3.2 Dificultad para separar tres tipos de plantas mediante los atributos *sepallength* y *sepalwidth* 26
3.3 Mejor separación de tres tipos de plantas mediante los atributos *petalwidth* y *sepalwidth* 27
3.4 Técnicas de aprendizaje 28
4.1 Modelo general de un proceso 33
4.2 Objetivos de un diseño experimental 35
6.1 Clasificación de Redes Complejas 48
6.2 Etapas en el desarrollo del proyecto 49
6.3 Generación de Redes Complejas 50
6.4 Estructura del vector de características 51
6.5 Diferencias significativas entre tipos de redes 52
6.6 Selección de características 54

6.7 Modelo del proceso de caracterización de redes complejas. 54
6.8 Clasificación de Redes Complejas 56
7.1 Gráficas Potencia vs Número de instancias. Modelo de efectos fijos. 62
7.2 Gráficas Potencia vs Número de instancias. Modelo mixto. 63
7.3 Histograma de la función $DDC(G)$ por tipo de red. 64
7.4 Histograma de la función $DDC(G)$ por número de nodos 65
7.5 Gráfica de los efectos principales para la función $DDC(G)$. 74
7.6 Gráfica de la interacción de factores para la función $DDC(G)$. 74
7.7 Gráfica de los residuales para la función $DDC(G)$ 76
7.8 Gráfica de los residuales para la función $DDC(G)$ *a*) por tipo de red, *b*) por número de nodos. 76
7.9 Árbol de decisión obtenido con el algoritmo C4.5 81
7.10 Identificación del tipo de red localmente 86
7.11 Tipo de red de los subgrafos de la instancia INT – 1997 86
7.12 Tipo de red de los subgrafos de la instancia INT – 2000 87
7.13 Tipo de red de los subgrafos de la instancia INT – 2003 87
b1. Histograma de la función $\langle k \rangle$ por tipo de red 96
b2. Histograma de la función $\sigma_{\langle k \rangle}$ por tipo de red 96
b3. Histograma de la función $L(G)$ por tipo de red 97
b4. Histograma de la función $D(G)$ por tipo de red 97
b5. Histograma de la función E_{loc} por tipo de red 97
b6. Histograma de la función E_{glob} por tipo de red 98
b7. Histograma de la función $CG(G)$ por tipo de red 98
b8. Histograma de la función $\langle k \rangle$ por número de nodos 100
b9. Histograma de la función $\sigma_{\langle k \rangle}$ por número de nodos 101
b10. Histograma de la función $L(G)$ por número de nodos 101
b11. Histograma de la función $D(G)$ por número de nodos 101
b12. Histograma de la función E_{loc} por número de nodos 102
b13. Histograma de la función E_{glob} por número de nodos 102
b14. Histograma de la función $CG(G)$ por número de nodos 102
d1. Gráfica de los efectos principales para la función $\langle k \rangle$ 111
d2. Grafica de la interacción de factores para la función $\langle k \rangle$ 111

d3. Gráfica de los efectos principales para la función $\sigma_{\langle k \rangle}$ 112

d4. Grafica de la interacción de factores para la función $\sigma_{\langle k \rangle}$ 112

d5. Gráfica de los efectos principales para la función $L(G)$ 113

d6. Grafica de la interacción de factores para la función $L(G)$ 113

d7. Gráfica de los efectos principales para la función $D(G)$ 113

d8. Grafica de la interacción de factores para la función $D(G)$ 114

d9. Gráfica de los efectos principales para la función E_{loc} 114

d10. Grafica de la interacción de factores para la función E_{loc} 115

d11. Gráfica de los efectos principales para la función E_{glob} 115

d12. Grafica de la interacción de factores para la función E_{glob} 115

d13. Grafica de los efectos de factores para la función $CG(G)$ 116

d14. Grafica de la interacción de factores para la función $CG(G)$ 116

e1. Gráfica de los residuales para la función $\langle k \rangle$ 118

e2. Gráfica de los residuales para la función $\langle k \rangle$ *a*) por tipo de red, *b*) por número de nodos 118

e3. Gráfica de los residuales para la función $\sigma_{\langle k \rangle}$ 119

e4. Gráfica de los residuales para la función $\sigma_{\langle k \rangle}$ *a*) por tipo de red, *b*) por número de nodos 119

e5. Gráfica de los residuales para la función $L(G)$ 120

e6. Gráfica de los residuales para la función $L(G)$ *a*) por tipo de red, *b*) por número de nodos 120

e7. Gráfica de los residuales para la función $D(G)$ 121

e8. Gráfica de los residuales para la función $D(G)$ *a*) por tipo de red, *b*) por número de nodos 121

e9. Gráfica de los residuales para la función E_{loc} 122

e10. Gráfica de los residuales para la función E_{loc} *a*) por tipo de red, *b*) por número de nodos 123

e11. Gráfica de los residuales para la función E_{glob} 123

e12. Gráfica de los residuales para la función E_{glob} *a*) por tipo de red, *b*) por número de nodos 124

e13. Gráfica de los residuales para la función $CG(G)$ 124

e14. Gráfica de los residuales para la función $CG(G)$ *a*) por tipo de red, *b*) por número de nodos 125

LISTA DE ECUACIONES

2.1 Distribución del grado 11

2.2 Coeficiente de agrupamiento local 11

2.3 Coeficiente de agrupamiento global 11

2.4 Longitud de ruta característica 12
2.5 Diámetro de la red 12
2.6 Eficiencia global 12
2.7 Eficiencia local 13
2.8 Coeficiente de dispersión del grado local 13
2.9 Desviación estándar del conjunto $\{i \cup \Gamma(i)\}$ 13
2.10 Grado promedio del conjunto $\{i \cup \Gamma(i)\}$ 13
2.11 Coeficiente de dispersión del grado global 13
2.12 Distribución del grado binomial 15
2.13 Distribución del grado Poisson 15
2.14 Distribución del grado Power-Law 16
2.15 Distribución del grado exponencial 18
2.16 Probabilidad de enlace preferencial, modelo BA 20
2.17 Probabilidad de enlace preferencial, modelo de Liu 21
4.1 Modelo de los efectos 36
4.2 Modelo de las medias 37
4.3 Modelo multivariado de las medias 39
4.4 Modelo multivariado de las medias, en forma matricial 39

LISTA DE CUADROS

7.1 Resultados del análisis de correlación de las funciones de caracterización 78
7.2 Comparación de algoritmos de clasificación que usan como entrada la función de caracterización *DDC*(*G*) 83
b1. Estadísticos descriptivos de las funciones de caracterización de acuerdo al tipo de red no importando el número de nodos 95
b2. Estadísticos descriptivos de las funciones de caracterización de acuerdo al tipo de red no importando el tipo de red 99
c1. Resultados MANOVA modelo de los efectos fijos 104
c2. Resultados MANOVA modelo mixto 106
f1. Resultados de la Prueba de Tukey para las funciones de caracterización 126

Capítulo 1

INTRODUCCIÓN

Vivimos en un mundo de redes, cualquier sistema complejo en la naturaleza o un sistema complejo artificial, como la estructura de una célula, las neuronas, la Web o Internet, pueden modelarse como una red compleja; donde los nodos o vértices son los elementos que conforman el sistema y las aristas o lados son las interacciones entre los elementos, la topología de una red compleja es el patrón que forman los nodos y las aristas de la red. Por ejemplo la estructura de una célula, se puede describir como una red de químicos conectados por reacciones químicas, el conjunto de páginas en la Web conectadas a través de sus enlaces, Internet modelado como un conjunto de ruteadores interconectados entre sí por medio de enlaces físicos [Latora 2001, Barabási 1999b].

Todos estos sistemas son difíciles de describir debido a su gran tamaño y la complejidad en la interacción entre sus elementos; para poder estudiarlos se describen como redes complejas [Barabási 1999a]. Las redes complejas comenzaron a ser estudiadas hacia finales de la década de los 50's principios de los 60's, mediante la Teoría de Grafos Aleatorios, pero en ausencia de redes reales de gran tamaño, esta teoría rara vez fue probada.

Pero esta teoría no permite describir con totalidad las características topológicas de las redes reales complejas, debido a los elementos que influyen en la conectividad de los

nodos, para ello es necesario el desarrollo de modelos que reproduzcan redes reales y funciones de caracterización que obtengan en términos cuantitativos las características topológicas que no pueden deducirse observando a la red compleja en general, para de esta manera comprender la dinámica y la estabilidad de las grandes redes complejas [Barabási 2003].

1.1 DEFINICIÓN DEL PROBLEMA

Muchas redes reales como la Internet, pueden ser modeladas como redes complejas. La búsqueda y navegación en redes complejas se ven afectados por los cambios que sufre la topología de la red.

Por esta razón es importante identificar características topológicas que permitan distinguir cuantitativamente en cada punto de la red su topología local, para que de esta manera procesos como la búsqueda puedan adaptarse a las características locales cambiantes y optimizar su desempeño.

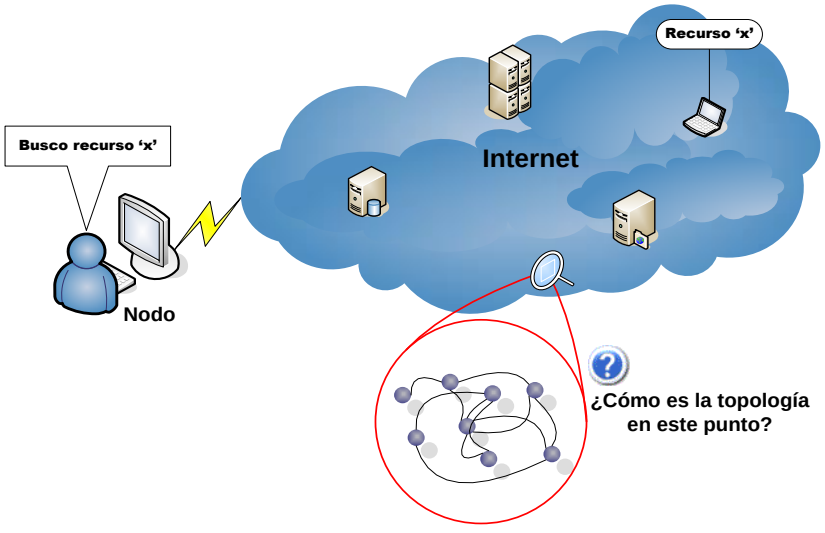


Figura 1.1 Planteamiento del problema

En la Figura 1.1, se ejemplifica el problema de la búsqueda de recursos. La nube representa la topología de Internet, que esta formada por computadoras y conexiones entre ellas. El círculo representa un subconjunto de nodos que forman parte de Internet. Así

cuando un nodo se conecta a la red en busca de un recurso (datos, otras computadoras) del cual no se sabe la ubicación, esta puede ser afectada por los cambios en la topología, el tráfico en la red, entre otros factores. En este tipo de situaciones es deseable contar con mecanismos para identificar la topología local, para que el proceso de búsqueda tome ventaja de las características topológicas locales a lo largo de su ejecución.

1.2 DEFINICIÓN FORMAL

Sea *C* un conjunto de clases que representan los tipos de redes complejas, *F* un conjunto de funciones de caracterización que se aplican sobre un grafo *G*, F' algún subconjunto de *F*, $\vec{f}$ un vector de valores que se obtiene al aplicar F sobre *G* y $\vec{f}'$ un vector de valores que se obtiene al aplicar F' sobre *G*.

Se desea encontrar el conjunto mínimo de funciones de caracterización F', tal que $P(C \mid F' = \vec{f}') @ P(C \mid F = \vec{f})$, donde $P(C \mid F' = \vec{f}')$ es la distribución de probabilidad de diferentes clases dados los valores de las características en F' y $P(C \mid F = \vec{f})$ es la distribución de probabilidad original dados los valores de las características en *F*.

Al conjunto mínimo de características que cumpla con $P(C \mid F' = \vec{f}') @ P(C \mid F = f)$ se le llama conjunto óptimo [Yu 2004].

1.3 JUSTIFICACIÓN

Una de las principales razones de estudio de las redes complejas es la flexibilidad y generalidad para representar de manera abstracta estructuras naturales, incluyendo los cambios dinámicos en su topología [Costa 2007]. Diversas investigaciones de sistemas complejos involucran la representación de la estructura de interés como una red compleja, seguido del análisis de características topológicas obtenidas de la representación realizada en términos de un conjunto de funciones de caracterización. Se han realizado estudios en los cuales se miden las propiedades estructurales de las redes complejas con el objetivo obtener una caracterización de cómo la conectividad de las redes complejas cambian con el tiempo y según los procesos que se efectúan en ellas. Ambas actividades, la representación

y la medición, tienen como objetivo la caracterización topológica de las estructuras estudiadas.

Por otro lado, se tiene especial interés en buscar funciones de caracterización que permitan discriminar entre los diferentes tipos de redes complejas. Tanto la caracterización y clasificación de estructuras naturales y artificiales modeladas como redes complejas implican una cuestión importante: ¿qué funciones de caracterización seleccionar de manera que permitan diferenciar los diferentes tipos de redes complejas? [Costa 2007].

La motivación principal de este trabajo es identificar funciones de caracterización que permitan discriminar diferentes tipos de redes complejas, para optimizar el desempeño de los procesos que se llevan a cabo en la red compleja, como los procesos de búsqueda de información, manejo de tráfico, diseño de algoritmos de ruteo, así como, el diseño de redes confiables. Todo esto con el fin de mejorar los sistemas tecnológicos existentes.

1.4 OBJETIVOS

Objetivo general

- Identificar un conjunto de funciones de caracterización que permitan discriminar entre diferentes topologías de redes complejas.

Objetivos particulares

- Implementar modelos de red que generan diferentes tipos de redes complejas.
- Implementar funciones de caracterización tales como: la distribución del grado, coeficiente de agrupamiento, longitud de ruta característica más corta, eficiencia de la red, coeficiente de dispersión del grado.
- Identificar las funciones de caracterización que proporcionen de manera cuantitativa información topológica que permita discriminar entre diferentes tipos de redes complejas.
- Identificar el clasificador con mejor desempeño.

1.5 HIPÓTESIS

Es posible identificar el conjunto mínimo de funciones de caracterización que permitan discriminar cuantitativamente entre diferentes topologías de redes complejas, mediante análisis estadístico y aprendizaje supervisado.

1.6 ORGANIZACIÓN DEL DOCUMENTO

El presente trabajo se encuentra estructurado de la siguiente manera, en el capítulo 2 se presentan fundamentos teóricos de redes complejas, en el capítulo 3 se revisan fundamentos teóricos aprendizaje automático, donde se hace énfasis en el aprendizaje supervisado y la selección de características, en el capítulo 4 se revisan los fundamentos teóricos del análisis y diseño de experimentos, en el capítulo 5 se presenta una revisión del estado del arte de la clasificación de redes complejas, en el capítulo 6 se presenta la metodología general de solución del problema y una metodología para seleccionar estadísticamente las características relevantes y no redundantes que permitan cuantitativamente identificar el tipo de una red, en el capítulo 7 se presenta la experimentación realizada y el análisis de resultados, en el capítulo 8 se encuentran las conclusiones y trabajos futuros.

Capítulo 2

REDES COMPLEJAS

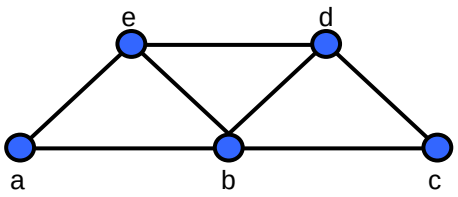
En este capítulo se abordan los fundamentos teóricos de redes complejas, estructurándose de la siguiente manera, en la sección 2.1 se revisan conceptos básicos de teoría de grafos los cuales permiten una mejor comprensión de las demás secciones, en la sección 2.2 se define que es un sistema complejo y como se modelan. Posteriormente en 2.3 se dan a conocer funciones de caracterización que permiten estudiar características topológicas de las redes complejas, en 2.4 se dan a conocer tipos de redes complejas y en 2.5 se presentan diferentes modelos para la generación de redes complejas, los cuales permiten reproducir grafos con características topológicas similares a las redes complejas reales.

2.1 TEORÍA DE GRAFOS

Cualquier sistema natural o artificial puede ser representado por una red compleja, una red compleja se refiere a un grafo con un gran número de elementos y que tiene una estructura topológica no trivial [Gusstafson 2006].

Un grafo es una abstracción matemática de situaciones donde se pueden representar elementos y sus relaciones. Un *grafo* se define como $G = (N, E)$, donde N es el conjunto de nodos y E es el conjunto de aristas, se representa matemáticamente mediante una matriz cuadrada $n \times n$, llamada *matriz de adyacencia M*, donde n es el número de nodos, cada

elemento de la matriz de adyacencia $M(i, j)$ representa la presencia o ausencia de una arista entre los nodos *i* y *j*. Si existe una arista entre los nodos *i* y *j* entonces $M(i, j) = 1$, en caso contrario $M(i, j) = 0$ [Latora 2001, Novaes 2005].



	a	b	c	d	e
a	0	1	0	0	1
b	1	0	1	1	1
c	0	1	0	1	0
d	0	1	1	0	1
e	1	1	0	1	0

N = (a,b,c,d,e)

E = {(a,b), (a,e), (b,c), (b,d), (b,e), (c,d), (d,e),}

Figura 2.1 Grafo y su correspondiente matriz de adyacencia

El número de nodos en el conjunto N_G es denominado *orden de un grafo* $n = |N_G|$ y el número total de aristas en el conjunto E_G es llamando *tamaño del grafo* $e = |E_G|$, el tamaño del grafo puede ser como mínimo 0 y como máximo $n(n-1)/2$. Tomando en consideración el grafo de la Figura 2.1, se tiene que el grafo es de orden 5 y tamaño 7 [Newman 2003, Albert 2002, Novaes 2005].

Dos nodos son *adyacentes* si una arista los une, el número de aristas adyacentes a un nodo es el *grado* del nodo *i* y se define como $k_i = \sum_{j=1}^{n} m_{ij}$, el *grado máximo* del grafo *G* se expresa como $\Delta(G)$, y el *grado mínimo* $\delta(G)$. El *grado promedio* del grafo *G* se representa como $\langle k \rangle = 2e/n$.

Los nodos adyacentes a un nodo *i* del grafo *G* se denominan *conjunto vecino* del nodo *i* y se denota por $\Gamma(i)$ [Latora 2001, Albert 2002, Newman 2003]. En la Figura 2.1 el grado de nodo B es $k_b = 4$, el conjunto vecino de b es $\Gamma(b) = \{a, c, d, e\}$ y el grado promedio del grafo $\langle k \rangle = (2*7)/5 = 2.8$.

Si dos nodos *i*, *j* no son adyacentes, pueden estar conectados por una secuencia de *m* aristas, al conjunto de aristas que forma la secuencia se le denomina ruta entre *i* y *j* siendo *m* la longitud de la ruta. A la ruta entre el nodo *i* y el nodo *j* con la mínima longitud posible, se le conoce como distancia geodésica d_{ij}, la cual puede ser calculada mediante la matriz de adyacencia.

Si *G* es conexo d_{ij} es positivo, lo que implica que el número de aristas para ir de *i* a *j* es finito. Por otro lado, si *G* no es conexo, puede ocurrir que no exista una ruta entre *i* y *j*, por lo cual $d_{ij} = \infty$. [Latora 2001].

Cuando se observa un grafo es difícil no enfocarse en como están arreglados los nodos y las aristas pero en la teoría de grafos lo que importa es el patrón que forman los nodos y sus aristas, es decir, la *topología* y no la figura que forman [Hayes 2000a]. En la Figura 2.2, se pueden observar tres grafos los cuales tienen diferente forma pero el patrón de conexión es el mismo.

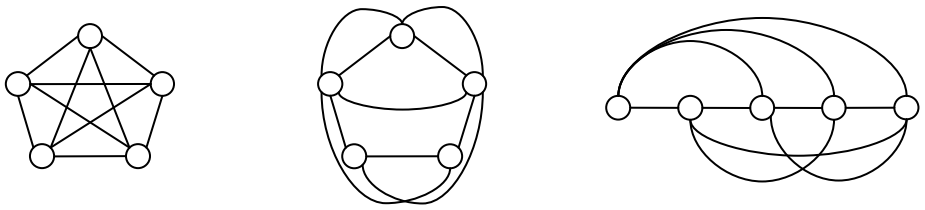


Figura 2.2 Grafos con la misma topología pero diferente forma

2.2 REDES COMPLEJAS

Un *sistema* es un conjunto de componentes relacionados entre si para lograr un fin común. Los diferentes sistemas naturales o artificiales pueden ser simples, complicados o complejos [Amaral 2004]. Los *sistemas simples* tienen un número pequeño de componentes los cuales actúan de acuerdo a leyes bien comprendidas. Los *sistemas complicados* tienen un gran número de componentes los cuales tienen roles bien definidos y están gobernados por reglas bien comprendidas. Los *sistemas complejos* tienen típicamente un gran número de componentes los cuales pueden actuar de acuerdo a reglas que pueden cambiar a través

del tiempo y que pueden no ser bien comprendidas, la conectividad de los componentes y sus roles son variables.

La física estadística trata con sistemas complejos, en donde predecir el comportamiento exacto de los componentes individuales podría ser imposible, por tanto, se limita a realizar predicciones estadísticas acerca del comportamiento colectivo de los componentes. Recientemente, se ha percibido que muchos sistemas formados por número muy grande de elementos pequeños obedecen leyes universales independientemente de los detalles microscópicos [Amaral 2001].

Debido a que cualquier sistema que pueda ser dividido en componentes relacionados por alguna regla, puede ser modelado como una red o grafo, en donde cada nodo representa un componente del sistema y una arista la interacción existente entre dos componentes cualesquiera; diversos estudios utilizan redes para modelar estos sistemas complejos dando lugar a las redes complejas.

Es en este punto donde la teoría de grafos y la física estadística se combinan para proveer las bases necesarias que permitan el estudio de las redes complejas a través de la *Teoría de Redes Complejas* [Costa 2007] y se enfoca en [Novaes 2005]:

1. Encontrar y destacar las propiedades estadísticas que caracterizan la estructura y el comportamiento de la red compleja, ofreciendo una forma adecuada de medir esas propiedades.
2. Crear modelos de redes que ayuden a entender el significado de esas propiedades.
3. Predecir el comportamiento de la red compleja basándose en el comportamiento de las propiedades estadísticas.

2.3 FUNCIONES DE CARACTERIZACIÓN

Las funciones de caracterización son medidas que tienen como objetivo proveer en términos cuantitativos las propiedades estructurales de la red; como se muestra en la Figura 2.3, al aplicar las funciones de caracterización sobre una red *G* se obtiene un vector de

características $\vec{f}_G$ [Costa 2007] que nos permiten caracterizar y analizar las redes complejas.

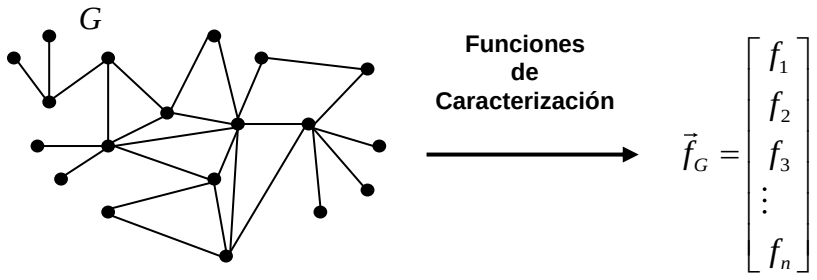


Figura 2.3 Extracción de características topológicas mediante funciones de caracterización

Las funciones de caracterización pueden clasificarse en dos tipos:

- Basadas en información global: requieren información de toda la red *G* para poder calcularlas.
- Basadas en información local: Considerando que cada nodo *i* en la red *G*, tiene asociado un subgrafo $G_i = (N_i, E_i)$, donde N_i es el conjunto $\Gamma(i)$, una función de caracterización local es una métrica que se calcula para cada G_i en *G*.

A continuación se explican algunas las funciones de caracterización.

2.3.1 Distribución del grado

El grado k_i de un nodo *i*, es una medida basada en información local. Definiremos $P(k)$ como la fracción de nodos en la red que tienen grado *k*, dicho de otro modo, $P(k)$ es la probabilidad de que un nodo seleccionado aleatoriamente tenga grado *k*. Una gráfica de $P(k)$ para cualquier red dada puede formarse haciendo un histograma de los grados de los nodos, este histograma es la distribución del grado de una red y esta basada en información global [Newman 2003].

Con base en lo anterior, definimos a X_k^G como el número de nodos en *G* con grado *k*. Así, la probabilidad de que un nodo en *G* tenga grado *k* se expresa por [Ortega 2005]:

$$P(k) = \frac{X_k^G}{n} \tag{2.1}$$

2.3.2 Coeficiente de agrupamiento

El coeficiente de agrupamiento mide la tendencia a formar grupos [Albert 2002, Newman 2003], para calcularlo se toma un nodo *i* de la red, con k_i aristas que lo conectan con k_i nodos. Si los vecinos cercanos del nodo original forman parte de un grupo, habría $k(k-1)/2$ aristas entre ellos. La proporción entre el número e_i de aristas que actualmente existen entre los k_i nodos y el total de $k(k-1)/2$ da el valor del coeficiente de agrupación para el nodo *i*, es $C(i)$.

$$C(i) = \frac{e_i}{\frac{k_i(k_i-1)}{2}} = \frac{2e_i}{k_i(k_i-1)} \tag{2.2}$$

El coeficiente de agrupamiento de toda la red $C(G)$ es el promedio del coeficiente de agrupamiento para cada nodo $i \in N$:

$$C(G) = \frac{\sum_{i \in N} C(i)}{n} \tag{2.3}$$

2.3.3 Longitud de ruta más corta característica

En la mayoría de los casos, dos nodos en una red compleja no son adyacentes, debido a que son redes esparcidas, sólo una fracción de todas las posibles aristas está presente. Aun así dos nodos no adyacentes *i* y *j* pueden estar conectados por una secuencia de *m* aristas, al conjunto de aristas que forma la secuencia se le denomina ruta entre *i* y *j* siendo *m* la longitud de la ruta.

Una característica que interesa estudiar es la longitud de ruta más corta que caracteriza a la red [Latora 2001, Albert 2002], para ello se define, la ruta geodésica (la más corta) como una de las rutas entre el nodo *i* y el nodo *j* con la mínima longitud posible, a

esta longitud se le conoce como distancia geodésica d_{ij}. De esta manera la longitud de ruta más corta característica $L(G)$, se define como el promedio de las distancias geodésicas.

$$L(G) = \frac{\sum_{i \geq j \in N} d_{ij}}{\frac{1}{2} n(n-1)} \equiv \frac{\sum_{i \neq j \in N} d_{ij}}{n(n-1)} \tag{2.4}$$

2.3.4 Diámetro de la red

El diámetro de la red *D*(*G*), se refiere a la longitud más grande de las distancias geodésicas. En las redes complejas el diámetro suele ser muy pequeño a pesar de la gran cantidad de nodos, a través de la matriz de las distancias geodésicas se puede calcular el diámetro.

$$D(G) = \underset{i \neq j \in N}{Max}(d_{ij}) \tag{2.5}$$

2.3.5 Eficiencia de la red

El concepto de eficiencia de la red *E*(*G*), es una medida de cómo la información es eficientemente intercambiada en toda la red [Latora 2001].

La eficiencia $\grave{o}_{ij}$ de comunicación entre los nodos *i* y *j* se define como $\grave{o}_{ij} = 1/d_{ij};\ \forall i, j$. Cuando no existe una ruta entre *i* y *j*, $d_{ij} = \infty$ y, consistentemente $\grave{o}_{ij} = 0$. La eficiencia promedio de *G* puede definirse como:

$$E(G) = \frac{\sum_{i \geq j \in N} \grave{o}_{ij}}{\frac{1}{2} n(n-1)} \equiv \frac{\sum_{i \neq j \in N} \grave{o}_{ij}}{n(n-1)} \tag{2.6}$$

La cantidad obtenida en la Ecuación (2.6), es la eficiencia global de *G* y la cual será referida como E_{glob}. También se puede caracterizar las propiedades locales de *G* evaluando para cada nodo *i*, de G_i, la eficiencia local el promedio de la eficiencia de los subgrafos locales:

$$E_{loc} = \frac{\sum_{i \in N} E(G_i)}{n} \tag{2.7}$$

La eficiencia local es similar al coeficiente de agrupamiento, debido a que manifiesta la tolerancia a fallas del sistema, mostrando que tan eficiente es la comunicación entre los vecinos del nodo *i*, cuando este es removido. Por otro lado la eficiencia global es una medida similar a la longitud de ruta característica pues muestra la eficiencia con la cual se comunican entre si dos nodos cualesquiera de la red.

2.3.6 Coeficiente de dispersión del grado

El coeficiente de dispersión del grado (*DDC*) [Ortega 2005] se define como una función basada en información local, la cual tiene como objetivo medir la dispersión entre el grado de un nodo *i* y $\Gamma(i)$. El *DDC* para el nodo *i* se define como:

$$DDC(i) = \frac{\sigma(i)}{\mu(i)} \tag{2.8}$$

donde,

$$\sigma(i) = \sqrt{\frac{\sum_{j \in \Gamma(i)} [k_j - \mu_i]^2 + [k_i - \mu_i]^2}{N_i}} \tag{2.9}$$

y,

$$\mu(i) = \frac{\sum_{j \in \Gamma(i)} [k_j] + k_i}{N_i} \tag{2.10}$$

representa la variación del grado entre el nodo *i* y $\Gamma(i)$, $\mu(i)$ es el grado promedio encontrado en el conjunto $\{i \cup \Gamma(i)\}$ y $N_i = |\{i \cup \Gamma(i)\}|$, además, k_i y k_j representan el grado de *i* y *j* respectivamente. Para calcular el coeficiente de dispersión del grado global, es decir para toda la red, se utiliza la siguiente fórmula:

$$DDC(G) = \frac{\sum_{i \in N} DDC(i)}{n} \tag{2.11}$$

donde *n* es el orden de la red.

2.4 TIPOS DE REDES COMPLEJAS

Las redes complejas, que representan sistemas complejos diferentes, comparten ciertas propiedades además de su gran tamaño, estas propiedades son [Hayes 2000b]:

- Tienden a ser esparcidas: tienden a tener relativamente pocas aristas en comparación con el gran número de nodos *n*, en general el número de aristas es más cercano a *n* que al número máximo de aristas que pueden existir.
- Tienen a ser agrupadas: las aristas en el grafo no están distribuidas uniformemente pero tienden a formar grupos.
- Tienden a tener diámetro pequeño: la ruta más larga de las cortas que atraviesan a una red compleja, puede estar alrededor de $\log n$, valor que es mucho mas chico que *n*.
- Tienen una secuencia de grado: conocida como distribución del grado, esta propiedad describe el patrón de conexión de los nodos.

De acuerdo a la distribución del grado, las redes complejas se clasifican en redes Aleatorias, redes Power-Law y redes Exponenciales, las cuales serán tratadas a continuación, así también, se explica el fenómeno "small world" que puede presentarse en estas redes.

2.4.1 Redes Aleatorias

La teoría de grafos se originó en el siglo XVIII con los trabajos de Leonard Euler acerca de la solución del problema de los puentes de Königsberg [Newman 2006]. En el siglo XX la teoría de grafos se volvió más estadística y algorítmica. Una particular fuente de ideas es el estudio de grafos aleatorios, grafos en los cuales las aristas se distribuyen aleatoriamente. [Barabási 1999a].

La teoría de grafos aleatorios fue publicada por Erdös y Rényi en una serie de artículos publicados a finales de los 50's y comienzos 60's. Erdös y Rényi se enfocaron a estudiar las propiedades estadísticas de grafos aleatorios con *n* nodos y *e* aristas. Erdös propuso el modelo de grafos aleatorios $G_{n,e}$ que considera a $M = \binom{n}{2}$ como el número

máximo de aristas posibles entre los *n* nodos del grafo *G*. Por lo anterior, se considera a $G_{n,e}$ como el espacio de todos los $\binom{M}{e}$ grafos posibles con *n* nodos y *e* aristas. [Bollobás 2002]

En 1959, Gilbert introdujo su modelo de grafos aleatorios $G_{n,p}$. La generación de grafos aleatorios se describe de la siguiente forma: Sea $\{X_{ij} : 1 \le i \le j \le n\}$ un arreglo de variables aleatorias de Bernoulli con $\Pr(X_{ij} = 1) = p$ y $\Pr(X_{ij} = 0) = 1 - p$, y sea $G_{n,p}$ un grafo con *n* nodos en el cual los nodos *i* y *j* son adyacentes, si $X_{ij} = 1$. En otras palabras, para construir un grafo aleatorio, $G_{n,p} \in G_{n,e}$, se agregan aristas con probabilidad independiente *p* [Bollobás 2002].

Más tarde a principios de los 80's diversas investigaciones arrojaron resultados más precisos, encontrando que un grafo aleatorio generado con una probabilidad de conexión *p* el grado k_i del nodo *i* sigue una distribución binomial con parámetros *n*-1 y *p*:

$$P(k) = \mathrm{C}_{n-1}^{k}\, p^{k}(1-p)^{n-1-k} = \binom{n-1}{k} p^{k}(1-p)^{n-1-k} \tag{2.12}$$

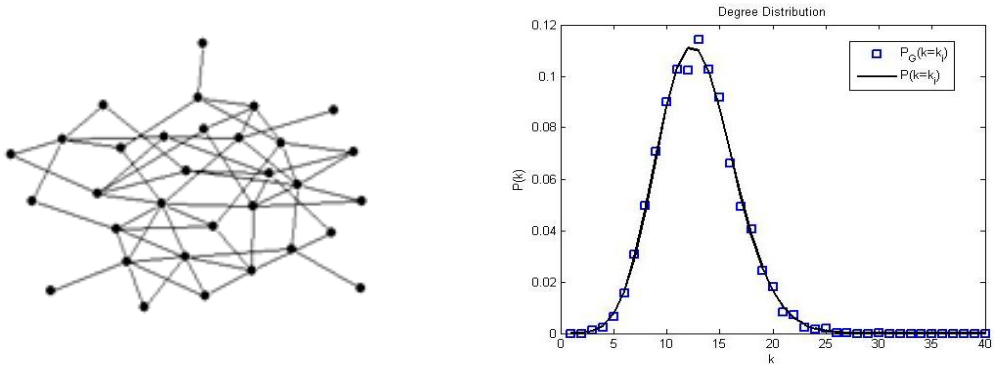
Para valores grandes de *n*, la distribución del grado sigue una distribución de Poisson, de ahí que la probabilidad de que un nodo tenga grado *k* es:

$$P(k) \sim \frac{\langle k \rangle^{k}\, e^{-\langle k \rangle}}{k!} \tag{2.13}$$

donde $\langle k \rangle = 2e/n = p(n-1)$ [Albert 2002, Barabási 2003, Newman 2003].

Dado que una red aleatoria las aristas entre los nodos son establecidas aleatoriamente, la mayoría de los nodos tiene aproximadamente el mismo grado, cercano al grado promedio de la red $\langle k \rangle$. En la Figura 2.4 a) se aprecia la estructura topológica de una red Aleatoria donde se puede observar como cada uno de los nodos se relaciona aproximadamente con la misma cantidad de nodos y en la Figura 2.4 b) se muestra la gráfica

distribución del grado obtenida para una red Aleatoria con 3072 nodos, 19778 aristas, $\delta(G)=3,\ \Delta(G)=30,\ \langle k\rangle=12$, los cuadros representan la distribución del grado real de la red y la línea continua representa la distribución del grado calculada con la Ecuación 2.13 y se observa como la distribución del grado sigue una distribución de Poisson.



a) b)

Figura 2.4 Red aleatoria: a) estructura topológica, b) gráfica de la distribución del grado.

2.4.2 Redes Power-Law

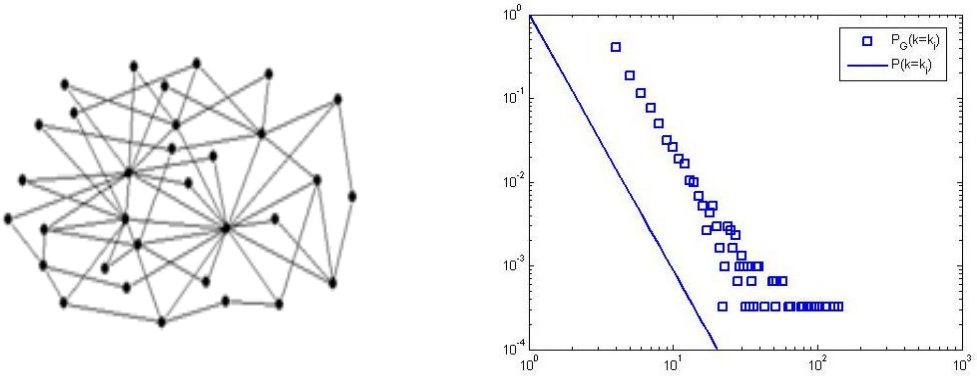
En la década de los 90's diversos investigadores como [Albert 2000, Faloutsos 1999], descubrieron que la distribución del grado en las redes del mundo real como el World Wide Web, Internet, redes de proteínas y metabolismo, las redes de lenguaje y sociales, difieren a la distribución de Poisson, exhibiendo una distribución del grado Power-Law:

$$P(k) \sim k^{-\gamma} \tag{2.14}$$

Las redes con distribución Power-Law son llamadas *Scale-Free* o *Libres de Escala*, debido a que independientemente de la escala, es decir del número de nodos, la distribución del grado no cambia. La característica invariante de estas redes es que un conjunto reducido de nodos que conforman la red tienen un gran número de enlaces (un grado muy alto) y resto de los nodos tienen pocos enlaces (un grado pequeño) [Barabási 2003, Newman 2003]. Esta característica permite que la tolerancia a fallas aleatorias sea grande, pero si los nodos con grado muy alto llamados nodos centrales son atacados intencionalmente esta red es muy vulnerable [Albert 2000].

En esta distribución el parámetro γ decrece desde ∞ hasta 0, siendo un parámetro de control que describe que tan rápido decae la frecuencia de aparición del grado *k*, de manera que el grado promedio de la red se incrementa a medida que γ se decrementa. En este caso $P(k)$ no tiene un pico definido, y para una *k* grande decae como una serie de potencias, apareciendo como una línea recta en una gráfica log-log.

En la Figura 2.5 a) se aprecia la estructura topológica de una red Power-Law donde se observa como unos cuantos nodos se relacionan con muchos nodos y el resto con muy pocos nodos y en la Figura 2.5 b) se muestra la gráfica distribución del grado obtenida para una red Power-Law con 3072 nodos, 9209 aristas, $\delta(G)=3$, $\Delta(G)=139$, $\langle k \rangle = 6$, los cuadros representan la distribución del grado real de la red y la línea continua representa la distribución del grado calculada con la Ecuación 2.14 y se observa como la distribución del grado sigue una distribución Power-Law.



a) b)

Figura 2.5 Red power-law: a) estructura topológica, b) gráfica de la distribución del grado.

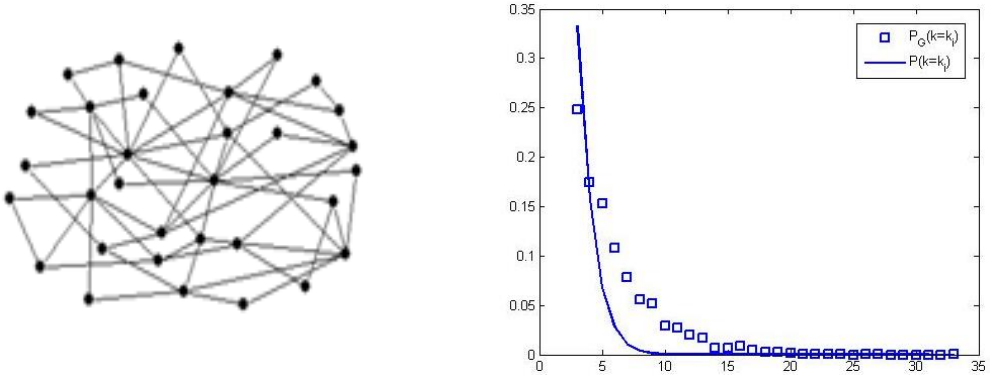
2.4.3 Exponenciales

Aunque las redes Power-Law son comunes, también existen casos de redes que exhiben una distribución del grado Exponencial como la red de energía del sur de California [Amaral 2000] y la red de vías de ferrocarril de la India [Sen 2003], a estas redes se les llama redes Exponenciales.

En estas redes la distribución del grado $P(k)$ muestra un pico en $\langle k \rangle$ y después cae exponencialmente para valores grandes de *k*:

$$P(k) \sim e^{-k} \tag{2.15}$$

Este tipo de red, no es muy tolerante a fallas aleatorias pues al eliminar cualquier nodo el daño es similar, debido a que cada nodo en la red se relaciona aproximadamente con la misma cantidad de nodos [Albert 2000]. En la Figura 2.6 a) se aprecia la estructura topológica de una red exponencial donde se observa que la mayoría de los nodos se relaciona aproximadamente con la misma cantidad de nodos y solo unos cuantos se relacionan con un mayor número de nodos, en la Figura 2.6 b) se muestra la gráfica distribución del grado obtenida para una red exponencial con 3072 nodos, 9211 aristas, $\delta(G)=3,\ \Delta(G)=33,\ \langle k \rangle = 6$, los cuadros representan la distribución del grado real de la red y la línea continua representa la distribución del grado calculada con la Ecuación 2.15 y se observa como la distribución del grado sigue una distribución Exponencial.



a) b)

Figura 2.6 Red exponencial: a) estructura topológica, b) gráfica de la distribución del grado.

2.4.4 Small World

El concepto "small world" es un término que describe el hecho de que a pesar de que las redes complejas tienen gran tamaño, en la mayoría de las redes la distancia entre dos nodos es relativamente corta. La distancia entre dos nodos esta definida como el número de nodos a lo largo de la ruta que los conecta.

La manifestación más popular de "small world" es el concepto de "seis grados de separación" descubierto por Stanley Milgram en 1967 quien concluyó que la ruta de conocidos entre pares de personas en los Estados Unidos tiene una distancia típica de seis. Esta propiedad de "small world" aparece caracterizando a la mayoría de las redes complejas. [Albert 2002]. El concepto "small world" no es indicador de un principio organizacional particular, sino un fenómeno que puede darse en las redes complejas.

2.5 MODELOS DE GENERACIÓN DE REDES

Los modelos de generación de redes, son una herramienta importante que reproducen redes que comparten las características topológicas de las redes del mundo real, el estudio de estos modelos nace con la finalidad de analizar y comprender funciones y fenómenos que se llevan a cabo en las redes complejas. De entre los modelos existentes podemos distinguir dos tipos de modelos, los modelos de generación sin crecimiento y modelos de generación basados en crecimiento que se describen en [Barabási 1999a].

2.5.1 Modelos de generación sin crecimiento

Estos modelos se caracterizan por dos aspectos importantes:

1. El número de nodos permanece constante, es decir, durante la construcción del grafo no se añaden nuevos nodos.
2. La agregación de aristas entre cualquier par de nodos sigue una distribución de probabilidad.

2.5.1.1 Modelo de Erdös y Rényi (ER)

El modelo de Erdös y Rényi [Albert 2002, Newman, 2003] es un modelo sin crecimiento que comienza con n nodos desconectados, cada par de nodos es conectado con una probabilidad p. Por tanto el número total de enlaces es un variable aleatoria con un valor esperado de $p[n(n-1)/2]$.

2.5.2 Modelos de generación basados en crecimiento:

Estos tipos de modelos se distinguen de los anteriores por los siguientes aspectos:

1. La construcción del grafo comienza con un número fijo de nodos y a cada paso cierta cantidad de nodos se añaden a la red, esto quiere decir que el número de nodos se incrementa.
2. La agregación de aristas entre cualquier par de nodos sigue una distribución de probabilidad.

2.5.2.1 Modelo Barabási-Albert (BA)

Después de haber estudiado el comportamiento de muchas redes reales hacia 1999, se introduce el modelo Barabási-Albert [Barabási 1999a] inspirado en el crecimiento y el enlace preferencial, este fue el primer modelo que reprodujo redes con una distribución del grado Power-Law. El modelo considera dos elementos:

- Crecimiento: comienza, con un pequeño número de nodos m_0, a cada paso *t* se añade un nuevo nodo con $m \leq m_0$ aristas que enlacen al nuevo nodo con m diferentes nodos ya existentes en el sistema.
- Enlace preferencial: cuando se selecciona los nodos a los cuales un nuevo nodo se va a conectar, se asume que la probabilidad Π de que un nuevo nodo sea conectado al nodo *i* depende del grado k_i del nodo *i*, tal que:

$$\Pi(k_i) = \frac{k_i}{\sum_{j \in N} k_j} \tag{2.16}$$

donde *N* es el conjunto de nodos en la red.

Después de *t* lapsos de tiempo, el procedimiento resulta en una red con $n = t + m_0$ nodos con *mt* aristas.

2.5.2.2 Modelo de Liu

Este modelo [Liu 2003] toma como referencia el modelo de Barabási, y propone un nuevo modelo basado en crecimiento considerando que en las redes reales en el enlace no es completamente preferencial, ni completamente aleatorio. De manera tal que la probabilidad Π_i de que un nuevo nodo sea conectado al nodo *i* existente en la red, debería contener

tanto un componente determinista que reflejen el enlace preferencial así como un componente aleatorio, por tanto se asume que:

$$\Pi_i = \frac{(1-p)k_i + p}{\sum_{j \in N}[(1-p)k_j + p]} \tag{2.17}$$

donde N es el número de nodos, $0 \leq p \leq 1$ es un parámetro que define el comportamiento determinista o aleatorio del modelo. Se puede observar que p es la probabilidad de que un nuevo nodo sea conectado aleatoriamente a un nodo i ya existente, y $(1 - p)$ es la probabilidad de que un nodo sea enlazado preferencialmente con el nodo i. Para $p = 0$ el modelo se reduce al de Barabási, y para $p = 1$ se elimina el enlace preferencial y se el establecimiento el enlace se vuelve completamente aleatorio, lo que permite crear redes con distribución Power-Law y Exponencial.

Capítulo 3

APRENDIZAJE AUTOMÁTICO

En este capítulo se abordan los temas de clasificación y selección de características, estos temas fueron fundamentales para el desarrollo de este trabajo, en la sección 3.1 se define que es un agente de aprendizaje, los componentes que lo conforman, en la sección 3.2 se aborda el tema de aprendizaje automático, se profundiza en el aprendizaje supervisado y la separabilidad de clases y una breve explicación de técnicas de aprendizaje utilizadas en aprendizaje automático, por ultimo la sección 3.3 aborda el tema de selección de características.

3.1 AGENTE DE APRENDIZAJE

Un *agente* es todo aquello que puede considerarse que percibe su ambiente mediante sensores y que responde o actúa en tal ambiente por medio de efectores, el *aprendizaje* es la idea de la que las percepciones deben servir no solo para actuar sino también para mejorar la capacidad de un agente para actuar en el futuro [Russell, 1996].

Los agentes con capacidad de aprender pueden ser divididos en cuatro componentes conceptuales como se puede apreciar en la Figura 3.1, estos componentes son:

- Elemento de aprendizaje: Toma parte del conocimiento acerca sí mismo y se retroalimenta con alguna información del comportamiento del agente, y decide

cómo hay que modificar el elemento de desempeño para que, de ser posible, en el futuro pueda estar en condiciones de mejorar su actuación.

- Elemento de desempeño: es el que escoge las acciones externas
- Critico: Su función consiste en informar al elemento de aprendizaje su evaluación del desempeño del agente.
- Generador de problemas: Tiene a su cargo proponer acciones que permitan obtener experiencias nuevas y que aporten información.

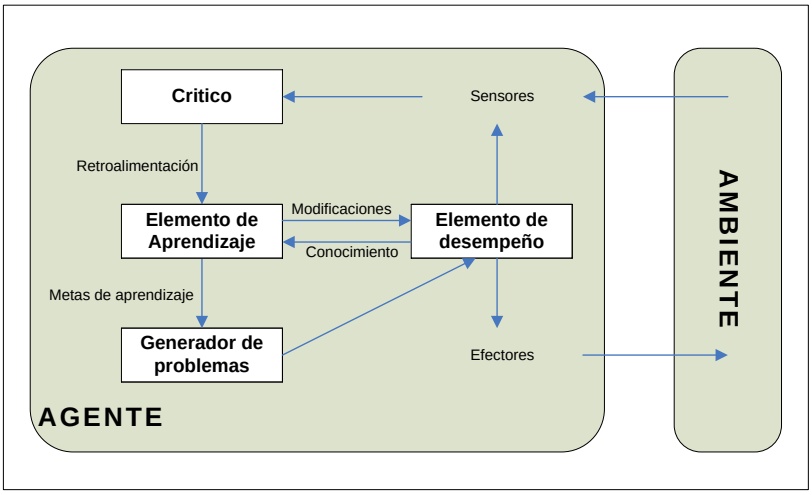


Figura 3.1 Arquitectura de un agente con capacidad de aprender

3.2 APRENDIZAJE AUTOMÁTICO

El aprendizaje automático es una rama de la inteligencia artificial que tiene como objetivo el estudio de técnicas que proveen, mediante el análisis de datos, capacidad de aprendizaje a un agente, es decir, que sea capaz de mejorar su rendimiento a través de la experiencia, que pueda adaptarse e integrar cierto conocimiento de manera que pueda resolver nuevos problemas [Mitchell 1997].

3.2.1 Tipos de aprendizaje

Dependiendo de la manera en que se da la retroalimentación podemos distinguir tres tipos de aprendizaje [Russell 1996]:

- Aprendizaje no supervisado: es el que se efectúa sin ninguna indicación sobre cuales son las salidas correctas. Desde el punto de vista de aprendizaje automático [Michie 1994] se tiene un conjunto de observaciones y el objetivo es determinar la existencia de clases o grupos en los datos; también se le conoce como clustering o agrupación.
- Aprendizaje por refuerzo: se da cuando el agente recibe una evaluación de sus acciones, sin informarle cuál habría sido la acción correcta. La evaluación ya sea negativa o positiva se le denomina *refuerzo*.
- Aprendizaje supervisado: el agente predice una determinada acción y el ambiente de inmediato ofrece una percepción que describe el resultado directo real. Desde el punto de vista de aprendizaje automático [Michie 1994] en este tipo de aprendizaje se conoce con certeza el número de clases o categorías existentes dentro de un conjunto de datos y el objetivo es establecer un modelo con el cual se puede clasificar una nueva observación en una de las clases o categorías existentes; a este tipo de aprendizaje también se le conoce como reconocimiento de patrones, discriminación o clasificación.

3.2.1.1 Aprendizaje Supervisado

En este tipo de aprendizaje conocido también como *aprendizaje de conceptos o inducción*, se adquiere conocimiento de ejemplos de entrenamiento, en donde cada ejemplo contiene características o atributos y una etiqueta que define una categoría o clase a la que pertenece, a esto se le llama clasificación del ejemplo; el objetivo es tomar un conjunto ejemplos e identificar con base en las características a que clase pertenecen [Michie 1994, Mitchell 1997].

De manera formal [Russell 1996] se dice que un ejemplo es un par $(x, f(x))$, en donde x es la entrada y $f(x)$ es la salida de la función que se aplico a x. El objetivo es inducir de la siguiente manera: dado un grupo de ejemplos de f producir una función h que aproxime a f. A la función h se le conoce como *hipótesis*.

Si un algoritmo de aprendizaje, produce una hipótesis o modelo que permita predecir satisfactoriamente las clases a las que pertenecen ejemplos que no han sido vistos anteriormente, se dice que el algoritmo de aprendizaje es bueno. Para evaluar la calidad de la

hipótesis se verifican sus predicciones en relación con la clasificación correcta una vez que se conoce, lo anterior se realiza en un conjunto de ejemplos conocidos como *conjunto de prueba*.

Se puede concluir que la clasificación consiste en la construcción de un modelo que será aplicado continuamente a una serie de ejemplos, en el cual cada nuevo ejemplo debe ser asignado a una de las clases predefinidas en base a características observadas.

Para evaluar la eficiencia de un algoritmo de aprendizaje, en este caso del clasificador, se recomienda seguir la siguiente metodología [Russell 2006]:

1. Reunir una cantidad considerable de ejemplos.
2. Dividirla en dos conjuntos disjuntos: conjunto de entrenamiento y de prueba.
3. Emplear el algoritmo de aprendizaje con el conjunto de entrenamiento, para producir una hipótesis *h*.
4. Medir el porcentaje de ejemplos del conjunto de prueba correctamente clasificados por *h*.
5. Repetir los pasos 1 a 4 en conjuntos de entrenamiento de tamaño diverso.

3.2.1.1.1 Separabilidad de clases

Las clases a las cuales pertenecen los ejemplos deben ser totalmente disjuntas, es decir, un ejemplo pertenece o no pertenece a una clase pero no puede pertenecer a dos clases a la vez [Mitchell 1997].

Si las clases son totalmente disjuntas, es decir, las clases son separables, el algoritmo de aprendizaje formará fronteras de clases, también llamadas fronteras de decisión, que pueden ser puntos, líneas, curvas, superficies e hipersuperficies dependiendo de la dimensionalidad del espacio características; es así, que el espacio de características se divide en segmentos, áreas, volúmenes o hipervolúmenes, llamados regiones de decisión, que por lo regular no se traslapan. [Kecman 2001, Duda 2000].

La manera en que el clasificador forma las fronteras de decisión tiene mucho que ver con la características que se tomen en cuenta para realizar la clasificación, para ilustrar esto, se tomo uno de los conjuntos de datos que podemos encontrar en Weka, `iris.arff`, este conjunto de datos contiene ejemplos de tres tipos de plantas (50 por cada tipo): *Iris-setosa*, *Iris-versicolor* e *Iris-viginica*. Cada ejemplo tiene cuatro características: el largo del sépalo (*sepallength*), el ancho del sépalo (*sepalwidth*), largo del pétalo (*petallength*) y ancho del pétalo (*petalwidth*).

Si se toman los atributos *sepalwidth* y *sepallength* como entradas de un clasificador BayesNet, se puede observar en la Figura 3.2, que las fronteras de decisión no logran separar adecuadamente las clases de plantas, ya que dentro de la región de decisión de la clase *Iris-versicolor* se pueden observar ejemplos de la clase *Iris-virginica* y viceversa.

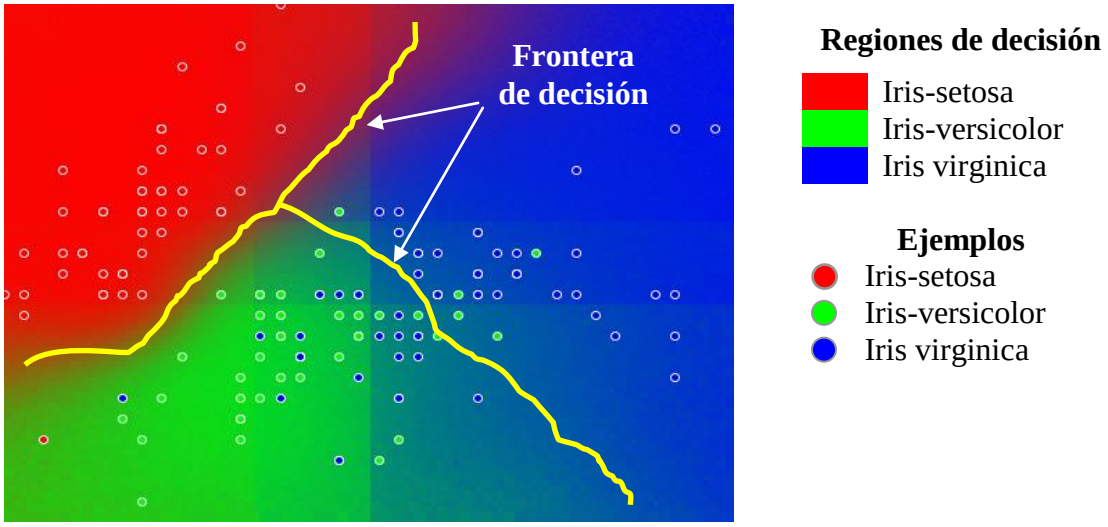


Figura 3.2 Dificultad para separar tres tipos de plantas mediante los atributos *sepallength* y *sepalwidth*

Por otra parte en la Figura 3.3 se muestra que tomando los atributos *petalwidth* y *sepalwidth* al mismo clasificador, este lograr definir de mejor manera las fronteras de decisión entre las clases, ya que se reduce el número de ejemplos ubicados fuera de sus regiones de

decisión. Estas dos figuras ilustran claramente la importancia de la selección de características, tema que se abordará en la sección 3.3.

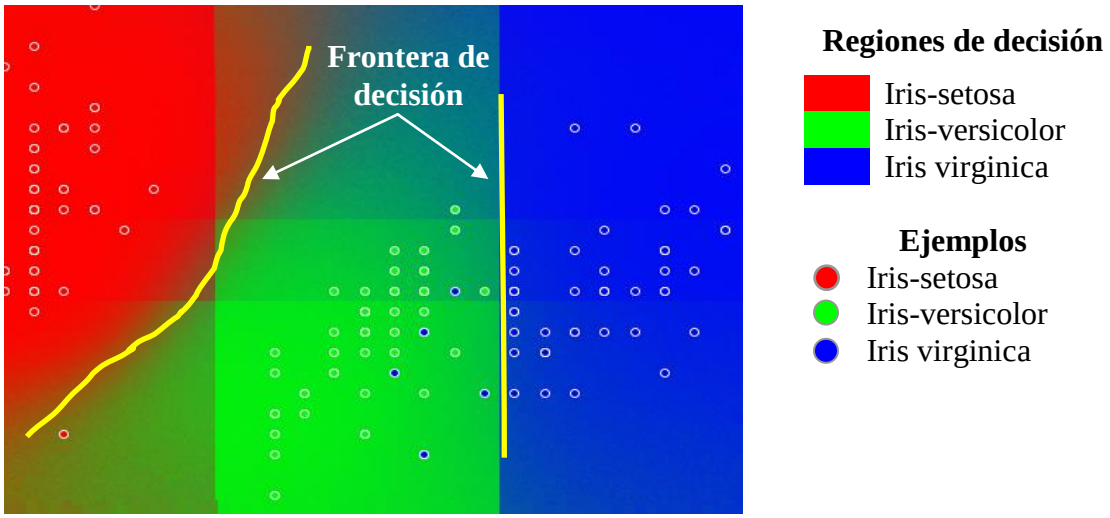


Figura 3.3 Mejor separación de tres tipos de plantas mediante los atributos *petalwidth* y *sepalwidth*

3.2.2 Técnicas de aprendizaje

Las técnicas de aprendizaje permiten encontrar y describir patrones estructurales en los datos, de manera que ayuden a explicar los datos y hacer predicciones a partir de estos; así también, las técnicas de aprendizaje buscan descripciones estructurales de lo que se ha aprendido, descripciones que se pueden convertir en algo complejo [Witten 2005], la Figura 3.4 muestra una clasificación de las técnicas de aprendizaje ampliamente usadas en el aprendizaje no supervisado y supervisado [Zhao 2007].

La clasificación es soportada por muchos métodos estadísticos, de aprendizaje automático y de redes neuronales. Algunos métodos ampliamente usados son: el Naive Bayes, la regresión logística y *k*-vecinos cercanos. Naive Bayes estima el radio de probabilidad bajo el supuesto de que las variables son condicionalmente independientes, la regresión logística encuentra una frontera lineal para separar dos clases, de esta manera se puede decir si un ejemplo pertenece o no a una clase. El *k*-vecinos cercanos simple memoriza los ejemplos de

entrenamiento y clasifica cada nuevo ejemplo dentro de la clase que tenga la mayor cantidad de los *k* ejemplos de entrenamiento cercano.

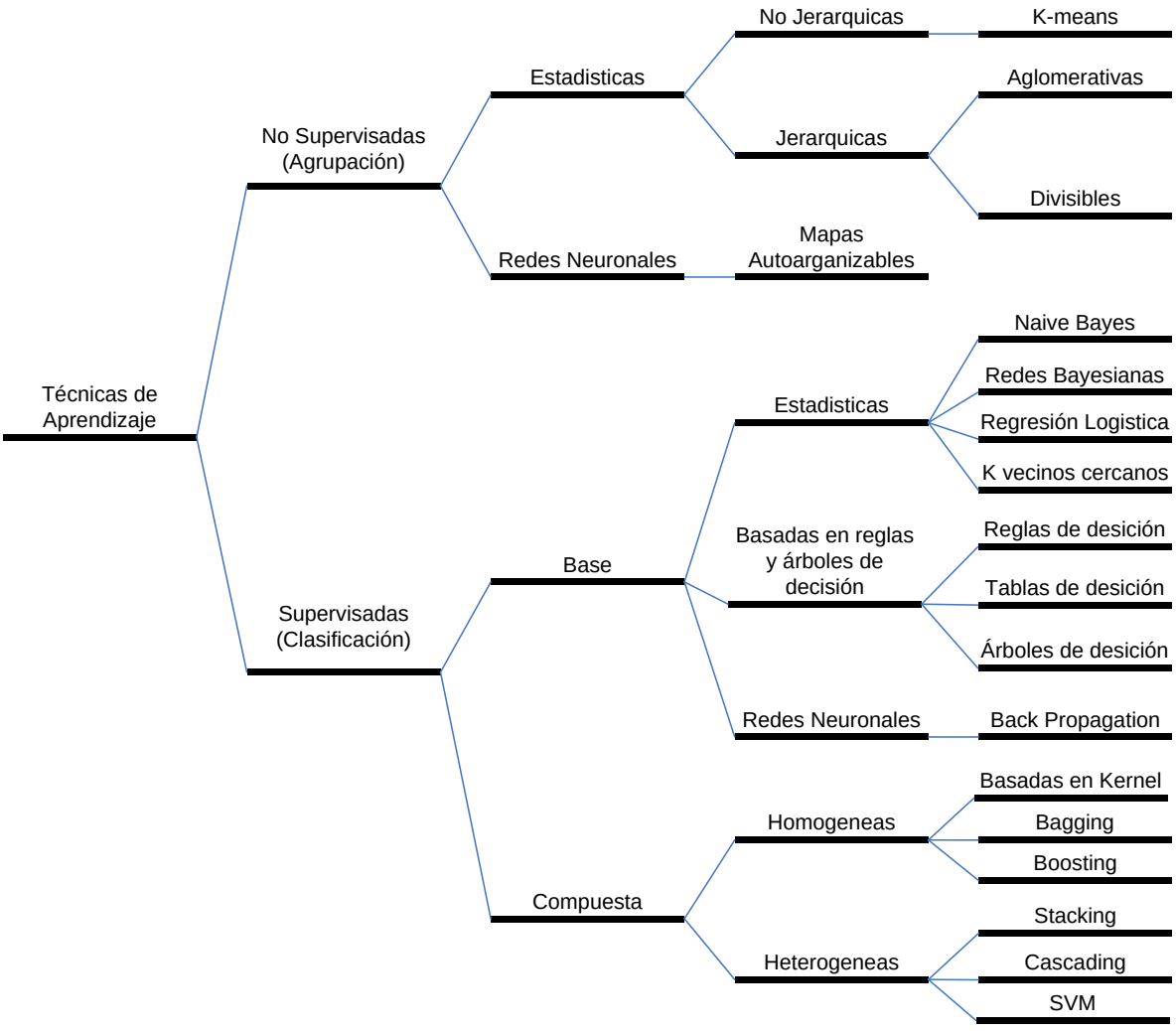


Figura 3.4 Técnicas de aprendizaje

Las técnicas de aprendizaje automático generan tablas, árboles y reglas de decisión. Backpropagation es una técnica de redes neuronales ampliamente usada para clasificar, las redes neuronales están altamente interconectadas con una capa de entrada, una capa de salida y cero o más capas intermedias, que ajustan sucesivamente los pesos de sus conexiones con los nodos de sus capas vecinas durante el entrenamiento.

Las técnicas de clasificación antes mencionadas se le conoce como técnicas base, algunas técnicas disponibles combinan múltiples clasificadores para mejorar la certeza de la clasificación estas técnicas se conocen como compuestas, ejemplos de estas son: bagging, boosting, cascading y stacking.

Bagging y boosting entrenan múltiples clasificadores (clasificadores base homogéneos) del mismo tipo haciendo una predicción final basada en los votos de los clasificadores base. En el método bagging los clasificadores base son entrenados independientemente usando diferentes conjuntos de entrenamiento y sus votaciones tienen igual peso. En boosting, los clasificadores aprenden secuencialmente, cada nuevo clasificador se enfoca más en los ejemplos mal clasificados por los clasificadores previos, el peso de su votación va de acuerdo a la certeza de cada clasificador.

Cascading y stacking combina clasificadores de diferentes tipos, cascading combina los clasificadores verticalmente con la salida de un clasificador base usada como variable de entrada adicional para el siguiente clasificador base. Stacking combina clasificadores horizontalmente, con las salidas de varios clasificadores usadas como variables de entrada para un clasificador de alto nivel responsable de hacer la decisión de clasificación final.

3.3 SELECCIÓN DE CARACTERÍSTICAS

En el área de clasificación la selección de características tiene grandes beneficios como [Guyon 2003]: facilitar la visualización y comprensión de los datos, reducir los requerimientos de recopilación y almacenamiento, reducir el tiempo de entrenamiento, y definitivamente mejorar el desempeño de los procedimientos de clasificación y construir modelos de clasificación simples y comprensibles.

En teoría, entre más características se consideren en el proceso de clasificación, estas deberían proveer mayor poder de discriminación, pero en la práctica, un conjunto de entrenamiento pequeño con una cantidad excesiva de características no sólo significa que el proceso de aprendizaje sea lento, si no que puede provocar un sobreajuste en el clasificador

con características que no proporcionan información debido a que son irrelevantes o redundantes con respecto al concepto de la clase, lo que afecta la exactitud del clasificador [Yu 2004, Witten 2005].

Se puede decir que el aprendizaje puede ser más eficiente y efectivo con características relevantes y no redundantes. Sin embargo el número posible de subconjuntos de características relevantes y no redundantes se incrementa exponencialmente al incrementarse la dimensionalidad de los datos, es decir, el número de características; por lo cual encontrar el subconjunto óptimo puede llegar a ser intratable y muchos problemas relacionados con la selección de características pueden ser vistos como un problema NP-duro [Singhi 2006, Guyon 2003].

Diversos investigadores han estudiados diversos aspectos de la selección de características, uno de los principales aspectos es medir la bondad de un subconjunto de características para determinar el óptimo. Podemos encontrar dos métodos de selección de características [Yu 2004, Witten 2005]:

- *Wrapper*: usa la exactitud predictiva de un algoritmo de aprendizaje predeterminado para determinar la bondad del conjunto seleccionado. Estos métodos son computacionalmente caros para datos con gran número de características.
- *Filter*: la selección subconjuntos de características son independientes de cualquier algoritmo de aprendizaje, se basa en medidas como la distancia, información, dependencia y consistencia de las características de los datos de entrenamiento.

La mayoría de los métodos para la selección de características involucra la búsqueda en el espacio de características del conjunto que ayuda a predecir mejor la clase, en cada etapa de la búsqueda un cambio local es realizado en el subconjunto de características, ya sea eliminando o añadiendo una característica, las estrategias de búsqueda pueden ser:

- Selección hacia delante: conocida como *forward selection*, comienza con un conjunto vacío, a cada paso una característica se añade tentativamente y el resultado del rendimiento del conjunto es evaluado, si el atributo añadido mejora el rendimiento del

conjunto la búsqueda continua de otro modo la búsqueda termina. Esta estrategia garantiza encontrar conjuntos localmente óptimos pero no necesariamente óptimos globales.

- Eliminación hacia atrás: conocida como *backward elimination*, comienza con el conjunto de todas las características, a cada paso se eliminan las características que no mejoran el desempeño del subconjunto, si todas las características mejoran el desempeño el procedimiento se detiene.
- Búsqueda bidireccional conocida como *bidirectional search*, combina los métodos de selección hacia delante y el de la eliminación hacia atrás. Este procedimiento selecciona características para añadirlas al conjunto, pero antes de la elección de otra característica se comprueba si las ya seleccionadas mejoran el desempeño, en algunos casos una característica puede ser útil al principio del proceso de selección, pero después de que se incluyen algunas características al conjunto puede ser ya no tal útil, con este procedimiento se eliminaría esta característica. El proceso se detiene cuando ninguna de las otras características a incluir mejoran el desempeño o cuando una característica a incluir es una de las que se acaban de eliminar.

Para evaluar las características, los métodos de selección explotan principalmente dos enfoques:

- Una característica a la vez: este método no puede capturar combinaciones que podrían dar buenos resultados, por lo que una característica que aparentemente no sirve por si sola, puede dar resultados muy buenos en combinación con otras, este método puede eliminar características irrelevantes, pero no las redundantes, ya que estas tienen una evaluación parecida a otras. Este enfoque es rápida y produce una lista ordenada de características de acuerdo a su medida de evaluación o ranking.
- Subconjuntos de características: este enfoque elimina la suposición de que las características son independientes entre si dada la clase, por lo que elimina tanto características irrelevantes como redundantes, produce una lista ordenada pero ahora de subconjuntos.

Capítulo 4

ANÁLISIS Y DISEÑO DE EXPERIMENTOS

En este capítulo se abordan los fundamentos teóricos de análisis y diseño de experimentos, estructurándose de la siguiente manera, en la sección 4.1 se explica que es un experimento, se presenta un esquema general de procedimientos para llevar a cabo una experimentación, y se explica un diseño de experimento, en la sección 4.2 se tratan brevemente métodos multivariados que se utilizan cuando existe más de una variable respuesta que se desea estudiar en el experimento. Posteriormente en 4.3 se trata el tema de la determinación del tamaño de la muestra, el cual es de vital importancia en cualquier diseño experimental.

4.1 ANÁLISIS Y DISEÑO DE EXPERIMENTOS

Para el desarrollo de la metodología se hizo uso del *análisis y diseño de experimentos*, que es una colección de herramientas estadísticas que se relacionan con la planeación, la ejecución y la interpretación de un experimento para obtener conclusiones validas y objetivas.

Un experimento puede definirse como una prueba planeada donde se introducen cambios controlados en las variables de un proceso o sistema con el fin de analizar cambios que pudieran observarse en las salidas del sistema, este proceso o sistema puede representarse con el modelo mostrado en la Figura 5.5 [Montgomery 2004].

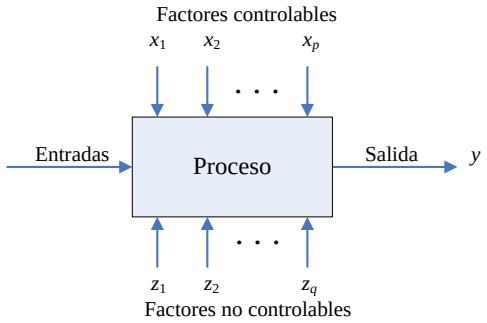


Figura 4.1 Modelo general de un proceso

Para aplicar el enfoque estadístico en un diseño y análisis de un experimento, es necesario contar con un esquema general de procedimientos a seguir. [Montgomery 2004] presenta una amplia explicación de este esquema, del cual se enuncian los procedimientos a seguir recomendados y que fueron adoptados en este trabajo:

1. Identificación y enunciación del problema: una enunciación clara del problema contribuye a una mejor comprensión de los aspectos bajo estudio y de la solución final.
2. Elección de factores, niveles y rangos: en este punto se deben considerar los factores que pueden influir en el sistema o proceso, cuales pueden ser controlados por el experimentador y cuales no; los niveles de los factores, es decir, valores específicos del factor, si son cualitativos o cuantitativos y el rango de valores que pueden tomar.
3. Selección de variables respuesta: identificar la variable o las variables que proporciona información útil acerca del proceso de estudio.
4. Elección del diseño experimental: implica la consideración del tamaño de la muestra, el número de factores y niveles, y el objetivo experimental.
5. Realización del experimento.
6. Análisis estadísticos de los datos: en esta fase deberán usarse métodos estadísticos para analizar los datos a fin de que las conclusiones sean objetivos y no de carácter apreciativo. Las técnicas estadísticas, aunadas con el conocimiento del proceso, llevan conclusiones sólidas.

7. Conclusiones y recomendaciones: es aquí donde se obtienen conclusiones prácticas acerca de los resultados y recomendar acciones a seguir.

4.1.1 Diseño Factorial

En muchos experimentos se tiene principal interés en estudiar los efectos o la influencia de dos o más factores sobre una variable de respuesta; para este tipo de experimentación, los diseños factoriales son los más eficientes. Casos especiales del diseño factorial general son usados ampliamente en trabajos de investigación debido a que constituyen las bases de otros diseños de gran valor práctico, como puede apreciarse en [Salazar 2005].

Por *diseño factorial* se entiende que en cada ensayo o réplica completa del experimento se investigan todas las combinaciones posibles de los valores de los factores llamados *niveles*. El *efecto* de un factor se define como el cambio en la variable de respuesta producido por un cambio en el nivel del factor, con frecuencia se le llama *efecto principal* porque se refiere a los factores de interés primario en el experimento. En algunos experimentos puede encontrarse que la diferencia en la respuesta entre los niveles de un factor no es la misma para todos los niveles de los otros factores, cuando esto ocurre existe una *interacción* entre los factores [Manson 2003, Montgomery 2004].

Hay situaciones en que las conclusiones se aplicarán únicamente a los niveles del factor considerado en el análisis, es decir, las conclusiones no pueden extenderse a niveles del factor que no se consideraron, entonces se dice que el factor es *fijo* y por lo regular se aplica cuando el factor tiene un número muy reducido de niveles. Existen situaciones en las que el factor tiene un gran número de posibles niveles y es deseable extender las conclusiones a la totalidad de niveles se hayan o no considerado en el análisis, entonces se dice que el factor es *aleatorio* [Montgomery 2004].

En general un diseño experimental, en este caso de un diseño factorial, tiene un conjunto de objetivos [Schmidt 1991] los cuales se ilustran en la Figura 4.2 y se enlistan a continuación:

a) Identificar factores que cambian la media de la variable respuesta pero no la variabilidad (*efecto de localización*, ver Figura 4.2a).

b) Identificar que factores contribuyen a la variabilidad de la variable respuesta pero no afectan la media (*efecto de dispersión*, ver Figura 4.2b).

c) Identificar que factores cambian la media y la variabilidad de la variable respuesta (ver Figura 4.2c).

d) Identificar que factores no tienen efecto alguno sobre la variable respuesta (ver Figura 4.2d).

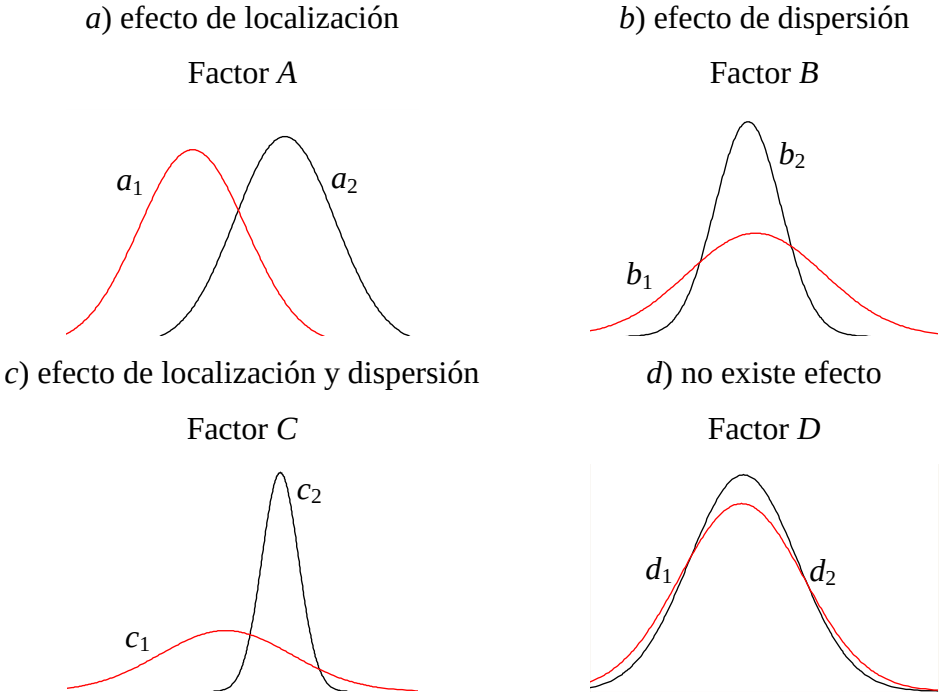


Figura 4.2 Objetivos de un diseño experimental

Los diseños factoriales ofrecen varias ventajas, son más eficientes que los experimentos de un factor a la vez, ya que puede haber interacciones entre los factores y evita llegar a conclusiones incorrectas; permiten la estimación de los efectos de un factor con varios niveles de los factores restantes, produciendo conclusiones que son válidas para un rango de condiciones experimentales.

Los tipos más simples de diseños factoriales incluyen dos factores. Hay *a* niveles del factor *A* y *b* niveles del factor *B*, los cuales se disponen en un diseño factorial; es decir, cada réplica del experimento contiene todas las *ab* combinaciones de los tratamientos. En general, hay *n* replicas.

En el caso general, sea y_{ijk} la variable respuesta observada cuando el factor *A* tiene el nivel *i*-ésimo (*i*=1,2,…,*a*) y el factor *B* tiene el nivel *j*-ésimo (*j*=1,2,…,*b*) en la réplica *k*-ésima (*k*=1,2,…,*n*), el experimento factorial de dos factores aparecerá como en la Tabla 4.1. En el orden en que se hacen las *abn* observaciones se selecciona al azar, por lo que este diseño es un diseño completamente aleatorizado.

Tabla 4.1 Arreglo general de un diseño factorial de dos factores

		Factor *B*			
		1	2	…	*B*
Factor *A*	1	$y_{111}, y_{112}, \cdots y_{11n}$	$y_{121}, y_{122}, \cdots y_{12n}$		$y_{1b1}, y_{1b2}, \cdots y_{1bn}$
	2	$y_{211}, y_{212}, \cdots y_{21n}$	$y_{221}, y_{222}, \cdots y_{22n}$		$y_{2b1}, y_{2b2}, \cdots y_{2bn}$
	⋮				
	a	$y_{a11}, y_{a12}, \cdots y_{a1n}$	$y_{a21}, y_{a22}, \cdots y_{a2n}$		$y_{ab1}, y_{ab2}, \cdots y_{abn}$

Las observaciones del un experimento factorial pueden describirse con un modelo. Hay varias formas de escribir el modelo de un experimento factorial. El *modelo de los efectos* es:

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n \end{cases} \tag{4.1}$$

donde μ representa el efecto global promedio de la variable respuesta, τ_i representa el efecto del nivel *i*-ésimo del factor *A*, β_j el efecto del nivel *j*-ésimo del factor *B*, $(\tau\beta)_{ij}$ es el efecto de la interacción entre el factor *A* y *B*, y ε_{ijk} es el error aleatorio.

Otro posible modelo de un experimento factorial es el *modelo de las medias*:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n \end{cases} \tag{4.2}$$

donde $\mu_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij}$, una descripción detallada del modelo de los efecto y el modelo de las medias puede encontrarse en [Manson, 2003][Montgomery, 2004] [Hinkelmann, 2005].

Cuando uno de los factores es fijo y el otro es aleatorio a este modelo se le llama análisis de la varianza del *modelo mixto*. De acuerdo con el tipo de los factores se plantean las hipótesis que se deben probar mediante el análisis de la varianza (ANOVA) de dos factores. En general las hipótesis se plantean partiendo de que los factores no afectan significativamente la media de las variables respuesta en los diferentes niveles de los factores, es decir, la media entre los diferentes niveles de los factores son significativamente iguales y los efectos de los factores y la interacción entre los factores sobre la media son cero, al realizar este análisis es importante que los datos cumplan con los supuestos de normalidad e independencia.

Cuando el análisis de la varianza muestra que las medias de las variables respuesta en los diferentes niveles de los factores difieren, por lo general es de interés hacer comparaciones entre las medias individuales de los niveles de los factores, para descubrir diferencias especificas, para ello existen diversas pruebas que nos ayudan a detectar estas diferencias como, la prueba de Tukey, el método LSD de Fisher, entre otras, una descripción detallada de estas pruebas y ejemplos pueden encontrarse en [Montgomery 2004].

4.2 MÉTODOS MULTIVARIADOS

Cuando en una experimentación no solo desea medir los efectos de uno o más factores sobre una sola variable respuesta, sino en múltiples variables respuestas, lo más apropiado es recurrir a los métodos multivariados.

Los métodos multivariados son útiles para analizar conjuntos grandes, complicados y complejos de datos que constan de una gran cantidad de variables respuesta medidas en números grandes de unidades experimentales [Jonhson 2000]. Los objetivos que persigue el análisis multivariado es resumir estos grandes conjuntos de datos por medio de relativamente pocos parámetros, y encontrar relaciones entre las variables respuestas, las unidades experimentales y ambas. Los métodos multivariados posen técnicas dirigidas a variables y dirigidas a individuos.

Las técnicas dirigidas a variables se enfocan en las relaciones que podrían existir entre las variables respuesta que se están midiendo, ejemplos de este tipo de técnica son el análisis de la matriz de correlación y el análisis de componentes principales. Las técnicas dirigidas por los individuos se enfocan en las relaciones que pondrían existir entre las unidades experimentales que se están midiendo, entre estas técnicas se pueden encontrar el, análisis multivariado de la varianza, el análisis por agrupación y el análisis discriminante [Jonhson 2000].

4.2.1 Análisis de la matriz de correlación.

Cuando dos variables no están correlacionadas se dice que no hay una relación lineal entre ellas, es decir, son variables independientes. El coeficiente de correlación proporciona una medida de la relación lineal entre dos variables continuas, cuando se analizan más de dos variables se obtiene una matriz de correlación en donde se muestra el coeficiente de correlación para cada par de variables. Este coeficiente puede variar en un rango de -1 a +1 y proporciona información acerca de los componentes de una correlación que son:

- La fuerza: entre más grande sea el valor absoluto del coeficiente más fuerte es la relación lineal entre las variables, un valor absoluto de 1 indica una perfecta relación lineal, el valor de 0 indica ausencia de relación lineal, en tanto, un valor intermedio es interpretado como una relación lineal débil.
- La dirección: indicada por el signo del coeficiente. Si ambas variables tienden a incrementarse o decrecer juntas, el coeficiente es positivo. Si una tiende a incrementarse y la otra a decrece, el coeficiente es negativo.

Es importante aclarar que el coeficiente de correlación solo mide relaciones lineales, esto quiere decir que una significativa relación no lineal puede existir aunque el coeficiente de correlación sea 0. Por esta razón es conveniente trazar una gráfica de dispersión de las variables, esta gráfica revelará si, en realidad, las variables están relacionadas entre si y como podrían estar relacionadas según la forma de la dispersión si tiene forma de línea, de curva o si los datos tienden a formar grupos, mediante esta gráfica también se puede observar la dirección y fuerza de la relación, así como detectar datos atípicos.

4.2.2 Análisis multivariado de la varianza.

Cuando se realiza un experimento de varios factores y en cada réplica del experimento hay p variables respuestas que se desean estudiar, el modelo de los efectos o de las medias, se extiende a un modelo multivariado de las medias, el cual se puede escribir como:

$$y_{ijkl} = \mu_{ijl} + \varepsilon_{ijkl} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n \\ l = 1, 2, \ldots, p \end{cases} \tag{4.3}$$

donde μ_{ijkl} representa el efecto global promedio de la variable *l*-ésima, en la *k*-ésima réplica dentro del nivel *j*-ésimo del factor *B* y del nivel *i*-ésimo del factor, y ε_{ijk} representa el error aleatorio. En forma matricial, ese modelo se puede escribir como: $\mathbf{y}_{ijk} = \boldsymbol{\mu}_{ij} + \boldsymbol{\varepsilon}_{ijk}$, donde:

$$\mathbf{y}_{ijk} = \begin{bmatrix} y_{ijk1} \\ y_{ijk2} \\ \vdots \\ y_{ijkp} \end{bmatrix}, \qquad \boldsymbol{\mu}_i = \begin{bmatrix} \mu_{ij1} \\ \mu_{ij2} \\ \vdots \\ \mu_{ijp} \end{bmatrix} \qquad \text{y} \qquad \boldsymbol{\varepsilon}_{ir} = \begin{bmatrix} \varepsilon_{ijk1} \\ \varepsilon_{ijk2} \\ \vdots \\ \varepsilon_{ijkp} \end{bmatrix} \tag{4.4}$$

Las hipótesis que se desean probar se plantean de manera similar a las que se plantean en el análisis de la varianza de dos factores, solo que considerando múltiples variables, lo que se conoce como análisis multivariado de la varianza (MANOVA), al igual que en el ANOVA se requiere que los datos analizados cumplan con los supuestos de normalidad.

Si el MANOVA indica que existen diferencias significativas entre los niveles de los factores, entonces se puede considerar las variables medidas una a la vez (ANOVA) para evaluar en donde ocurren en realidad las diferencias entre los niveles de esos factores.

4.2.3 Análisis discriminante

Supóngase que se tienen unidades experimentales que provienen de diversos grupos o clases, pero no se sabe de cual de esos grupos pertenecen las unidades experimentales, el análisis discriminante tiene como objetivo predecir a que clase pertenecen las unidades experimentales con base a un conjunto de variables predictoras, para ello produce reglas discriminantes o esquemas de clasificación que nos indican a que clase es más probable que pertenezca una unidad experimental.

Para desarrollar una regla discriminante que clasifique las unidades experimentales en una de las varias clases, se debe tener una muestra aleatoria de unidades experimentales de cada clase. Algebraicamente la regla discriminante representa una combinación lineal de variables predictivas de manera que se maximice la diferencia entre los valores promedios de las clases, dicho de otra manera, la combinación lineal busca maximizar la variabilidad entre las clases, minimizando la variabilidad dentro de las clases.

4.3 DETERMINACIÓN DEL TAMAÑO DE LA MUESTRA.

Pueden cometerse dos tipos de errores cuando se prueban hipótesis. Si la hipótesis nula se rechaza cuando es verdadera, ha ocurrido un error de tipo I, la probabilidad de que este error ocurra se representa con α, conocida como *nivel de significación*. Si la hipótesis no se rechaza cuando es falsa, se ha cometido un error de tipo II, la probabilidad de que este error ocurra se representa con β y es conveniente que sea un valor pequeño.

El tamaño de la muestra y la probabilidad de ocurra un error de tipo II, guardan una estrecha relación, por esta razón, en cualquier problema de diseño experimental, una decisión crítica es la elección del tamaño de la muestra, es decir, el número de réplicas que deben efectuarse del experimento.

En general, si se tiene interés en detectar efectos pequeños de los factores en las variables respuestas, se necesitan más replicas que cuando se interesa en detectar efectos grandes. Para seleccionar el número de replicas a efectuar de manera que el diseño experimental sea sensible a diferencias potenciales importantes en los niveles de los factores una recomendación es hacer uso de las curvas de operación características.

Una curva de operación característica es una gráfica de probabilidad del error tipo II de una prueba estadística para un tamaño de muestra particular contra un parámetro que refleja la medida en que la hipótesis nula es falsa. En [Montgomery, 2004] se muestran diversas formulas derivadas de las curvas de operación características para determinar el número de replicas, dependiendo del diseño experimental seleccionado.

Capítulo 5

ESTADO DEL ARTE

En este capítulo se realiza una reseña de trabajos de investigación relacionados con la clasificación de redes complejas mediante características topológicas. En la sección 4.1 se describe como se ha abordado este tema, en campo de la ciencia, que modelos de generación de redes, funciones de caracterización y técnicas de aprendizaje se ha empleado y sus resultados. En la sección 4.2 se realiza una discusión de los trabajos relacionados y una comparación entre los trabajos relacionados con el trabajo propuesto en esta tesis.

5.1 TRABAJOS RELACIONADOS

En [Ali 2004] el objetivo es clasificar redes de comunicación dentro de tres tipos de topología, los investigadores no muestran evidencias de haber usado redes reales. Desarrollaron un sistema de reconocimiento para clasificar entradas de patrones topológicos de redes regulares, aleatorias y power-law. Para llevar a cabo dicho trabajo se empleó una red neuronal autoorganizable que usa como entradas los eigenvalores de la matriz de adyacencia de una red de comunicación. El conjunto de entrenamiento que utilizan es de 300 patrones por cada tipo de red; el tamaño de cada red fue de 56 nodos, no se reporta el número de aristas de las redes. Los autores concluyen que las entradas de los patrones se clasifican exitosamente pero no reportan datos que soporten sus conclusiones.

En [Middendorf 2004], el objetivo es determinar de un conjunto de modelos aquellos que describen más adecuadamente un conjunto de redes que representan sistemas biológicos. El conjunto de redes reales utilizado en este trabajo fueron: la red genética de la bacteria E. coli (423 nodos, 519 aristas), la red neuronal del gusano C. elegans (306 nodos, 2359 aristas) y la red de interacción de la levadura S. cerevisaie (2214 nodos, 2203 aristas). Utilizan 17 modelos para generar redes dirigidas y no dirigidas. Utilizan una técnica que realiza operaciones sobre la matriz de adyacencia para extraer características que sirven como entradas al clasificador, esta técnica es descrita en [Ziv 2005]. El algoritmo de clasificación usado con éxito fue el SVM con kernel lineal, se utilizaron 1000 instancias por modelo para cada conjunto de datos reales. Según los resultados reportados los modelos que mejor describen a los conjunto de datos reales, son modelos que se basan en mecanismos de duplicación y mutación. Si bien los resultados de clasificación obtenidos son buenos, no muestran si las características extraídas mediante la técnica descrita en [Ziv 2005] mejoran el desempeño del clasificador con respecto a información proporcionada por funciones de caracterización ampliamente usadas en el campo de las redes complejas.

Por otra parte, [Airoldi 2005] presenta una experimentación con el objetivo de valorar la estabilidad y separabilidad de diferentes tipos de redes. Como entradas a los clasificadores utilizan 47 funciones de caracterización. Toma en cuenta seis topologías de redes: de anillo, small world, aleatorias, core-periphery, power-law, celulares, estas redes se generaron con modelos disponibles en el proyecto ORA [CASOS 2005]. Utilizaron clasificadores como: Naive Bayes, Regresión Logística, Máxima Entropía, SVM con kernel lineal, Perceptron, Árboles de decisión y el *k*-vecino cercanos, siendo Naive Bayes el reportado con mejor desempeño. Los resultados muestran que usando el conjunto de funciones de caracterización se pudo distinguir casi exactamente para las redes aleatorias, power-law que modelo se utilizó para generar estas topologías, el porcentaje de error reportado para las redes aleatorias son del 0% y para las redes power-law el 0.07%. Para las redes celulares se distinguió que modelo las generó con un porcentaje de error del 17.64%. Para las redes “small world” se distinguió pobremente que modelo generó con un porcentaje de error del 24.78% y para las redes core-periphery no se pudo distinguir que modelo generó las redes, el porcentaje de error observado es el 50%.

En el trabajo reportado en [Costa 2007], se presenta una posibilidad para clasificar redes complejas combinando el análisis canónico de variables y la teoría de decisión bayesiana. El objetivo es clasificar redes complejas experimentales en tres categorías definas por tres modelos: BA, ER y Geographical Network (GN). Las redes experimentales que se consideraron fueron: la red de transporte de las aerolíneas de Estados Unidos (USATN), la red de interacción de proteínas de S. cerevisiae, los sistemas autónomos de Internet del año 1998, red de genes de E. coli, la red Delaunay. Un total de tres conjuntos de 300 instancias de red por modelo fueron generados con un grado promedio cercano al grado promedio de las redes experimentales, de las redes experimentales y las generadas se extrajeron 9 funciones de caracterización. Se realizaron 8 diferentes combinaciones con las funciones de caracterización y se aplicó un análisis canónico para reducir la dimensionalidad de los datos y posteriormente se aplica el método de decisión, los resultados muestran que el potencial del procedimiento de clasificación para identificar la clase de un red de naturaleza desconocida, varía de acuerdo a la combinación de funciones de caracterización tomadas al azar, debido a que la presencia de propiedades topológicas especificas en algunas redes experimentales no son totalmente compatibles con los modelos teóricos, además se puntualiza que un excesivo número de características puede comprometer la calidad de la clasificación.

5.2 DISCUSIÓN DE LOS TRABAJOS RELACIONADOS

Como se puede observar en la sección anterior diversos investigadores como [Ali 2004, Middendorf 2004, Airoldi, 2005] tienen como objetivo identificar el tipo de red compleja al que pertenece una red o asociar los grafos generados mediante algún modelo con una red del mundo real, para ello utilizan técnicas de aprendizaje supervisado.

Similar a los trabajos mencionados en la sección 5.1, este trabajo tiene como objetivo identificar el tipo de red mediante funciones de caracterización, sin embargo, se propone realizar una selección de características mediante análisis estadístico que permita mejorar el desempeño de las técnicas de aprendizaje supervisado. Es importante mencionar que los trabajos relacionados no realizan una selección de características relevantes y no redundantes que pudieran mejorar el desempeño de las técnicas de aprendizaje utilizadas.

Otro punto importante de comentar es la técnica de aprendizaje aplicadas son diversas y según lo que concluye cada uno de los investigadores los resultados de la red neuronal, el SVM con kernel lineal, Naive Bayes resultaron las mejores técnicas para discriminar redes complejas. Por lo tanto se tomarán como base para medir el poder discriminante de las características seleccionadas, para medir el desempeño de las técnicas de aprendizaje con las características se utilizará el programa WEKA.

También se pudo observar que la naturaleza de los datos es diversa, esto nos indica que el estudio de las redes complejas es multidisciplinario, también se concluye que para validar los resultados es importante el estudio de instancias reales. Para este trabajo se utilizarán instancias de Internet que reflejan los cambios de su topología en el tiempo.

5.2.1 Análisis Comparativo

En la Tabla 5.1, se presenta un análisis comparativo de los trabajos relacionados con la clasificación de redes complejas usando funciones de caracterización que discriminen entre diferentes tipos de redes.

Tabla 5.1 Análisis comparativo de los trabajos relacionados

Trabajo Relacionado	Análisis estadístico	Selección de características	Funciones de caracterización	Instancias de redes reales
[Alí 2004]			✓	
[Middendorf 2004]				✓
[Airoldí 2005]			✓	✓
[Costa 2007]	✓		✓	✓
Propuesta de tesis	✓	✓	✓	✓

En la columna 2 indica si se realiza un análisis estadístico de las funciones de caracterización, la columna 3 muestra si se lleva a cabo una selección de funciones de caracterización que representen el conjunto mínimo de manera que se mejore el desempeño de las técnicas de aprendizaje, en la columna 4 se indica si se utilizaron funciones de caracterización usadas en el campo de las redes complejas, en la columna 5 se muestra si

trabajan con instancias de redes que representan sistemas del mundo real, las cuales están disponibles en diversos sitios en Internet (ver Anexo A).

En contraste con los trabajos previos, este trabajo de tesis considera los cuatro puntos de la Tabla 5.1, los cuales son relevantes para el problema de la clasificación de redes complejas.

Capítulo 6

METODOLOGÍA

Diversos investigadores [Alí 2004, Middendorrf 2004, Airoldi, 2005], se han dado a la tarea de discriminar entre diferentes tipos de redes complejas o relacionar instancias generadas mediante algún modelo con una red compleja del mundo real, utilizando algoritmos de clasificación. Como entradas al algoritmo utilizan la matriz de adyacencia o funciones de caracterización, sin embargo no muestran suficientes evidencias de un análisis detallado de las funciones de caracterización que permita determinar sí son el conjunto mínimo para realizar una discriminación eficaz o cuáles son las mejores para discriminar.

En este capítulo se presenta la metodología empleada para identificar un conjunto de funciones de caracterización que permitan discriminar entre diferentes topologías de redes complejas. El capítulo está organizado como sigue: en la sección 6.1 se describe la metodología propuesta y las tres etapas principales en el desarrollo de este trabajo de tesis, en las secciones 6.2, 6.3 y 6.4 se describen detalladamente las actividades realizadas en cada etapa de este trabajo.

6.1 METODOLOGÍA PROPUESTA

Esta metodología toma como base la arquitectura de un agente de aprendizaje, ilustrada en la Figura 3.1 y tiene como objetivo identificar el conjunto mínimo de funciones de caracterización que permitan discriminar entre diferentes topologías de redes complejas.

En esta metodología se toma como *ambiente* un conjunto de redes complejas caracterizadas de las cuales se conoce su tipo, sobre estas instancias se realiza una selección de funciones de caracterización que permitan mejorar el desempeño de las técnicas de aprendizaje supervisado (clasificador). En el *entrenamiento* se mide el desempeño del clasificador entrenando al agente y se construye un modelo. En la *predicción* se clasifican nuevas redes de tipo desconocido mediante el modelo construido, estas redes están caracterizadas con las funciones de caracterización que pertenecen al conjunto mínimo, esto se puede apreciar en la Figura 6.1.

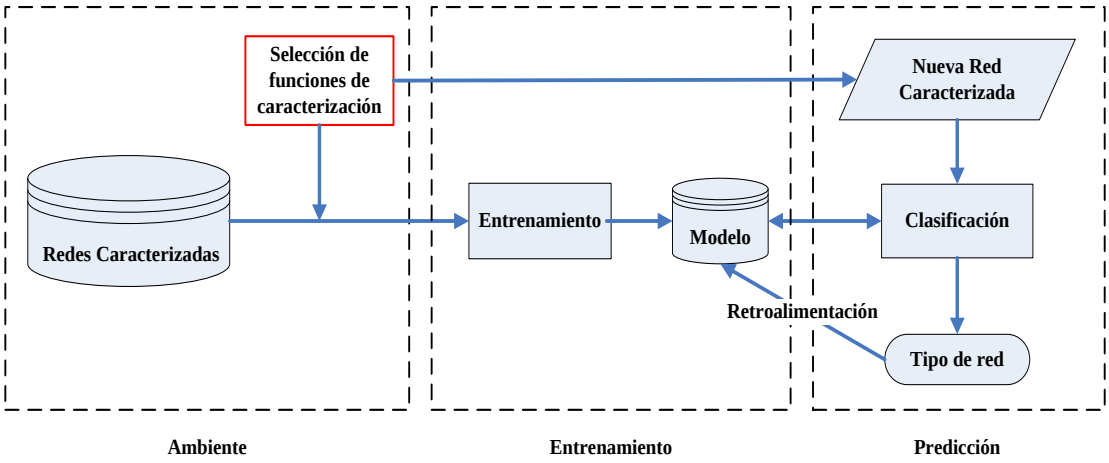


Figura 6.1 Clasificación de Redes Complejas

En la Tabla 6.1 se presenta una analogía entre la arquitectura de un agente de aprendizaje y la clasificación de redes complejas.

Tabla 6.1 Arquitectura de un agente de aprendizaje vs. Clasificación de redes complejas

Agente de aprendizaje	Clasificación de Redes Complejas
Generador de problemas y metas de aprendizaje	Diferentes tipos de redes caracterizadas
Elemento aprendizaje, Modificaciones	Entrenamiento
Conocimiento	Modelo
Retroalimentación	Retroalimentación
Elemento de desempeño	Clasificación
Sensores	Nueva red caracterizada

Continuación Tabla 6.2 Arquitectura de un agente de aprendizaje vs. Clasificación de redes complejas

Efectores	Tipo de red
Ambiente	Redes complejas

Las actividades llevadas a cabo para lograr el objetivo planteado se pueden dividir en tres etapas que pueden apreciarse en la **Figura 6.2**. La primera etapa es la *generación de redes complejas*, el producto de esta etapa es un conjunto de instancias de red caracterizadas *I*; la segunda etapa es *selección de características* donde se realiza un análisis estadístico de las instancias de red para determinar las funciones de caracterización que mejor discrimina entre los tipos de redes y la tercera etapa es la *clasificación de redes complejas* caracterizadas con las funciones seleccionadas y la determinación del clasificador con el mejor desempeño.

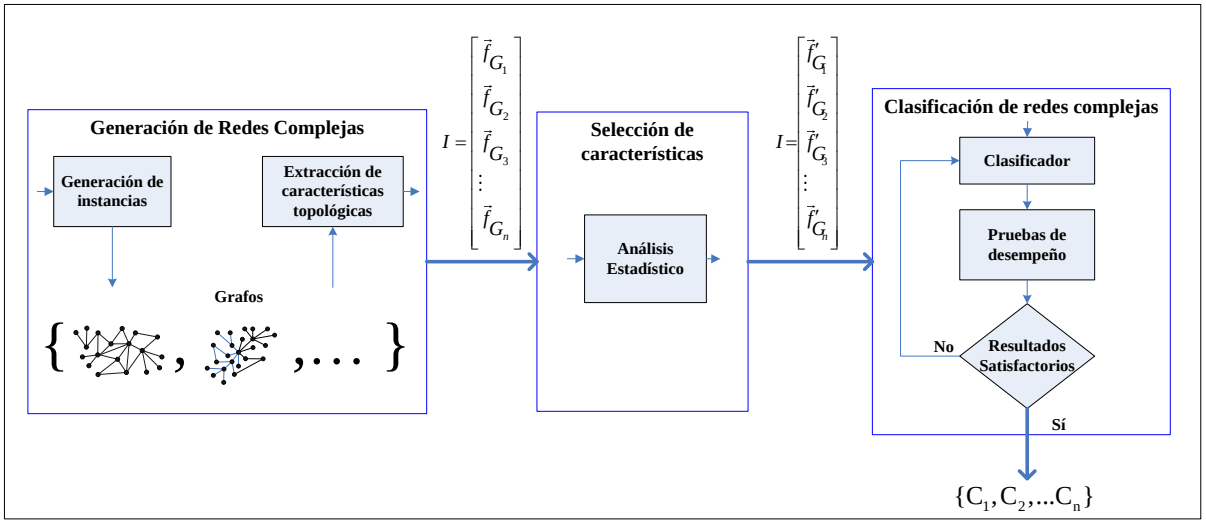


Figura 6.2 Etapas en el desarrollo del proyecto

6.2 GENERACIÓN DE REDES COMPLEJAS

En esta etapa se realizaron dos actividades, como puede apreciarse

Figura 6.3, la generación de instancias de red y la extracción de características topológicas mediante las funciones de caracterización, estas dos actividades se describirán detalladamente a continuación.

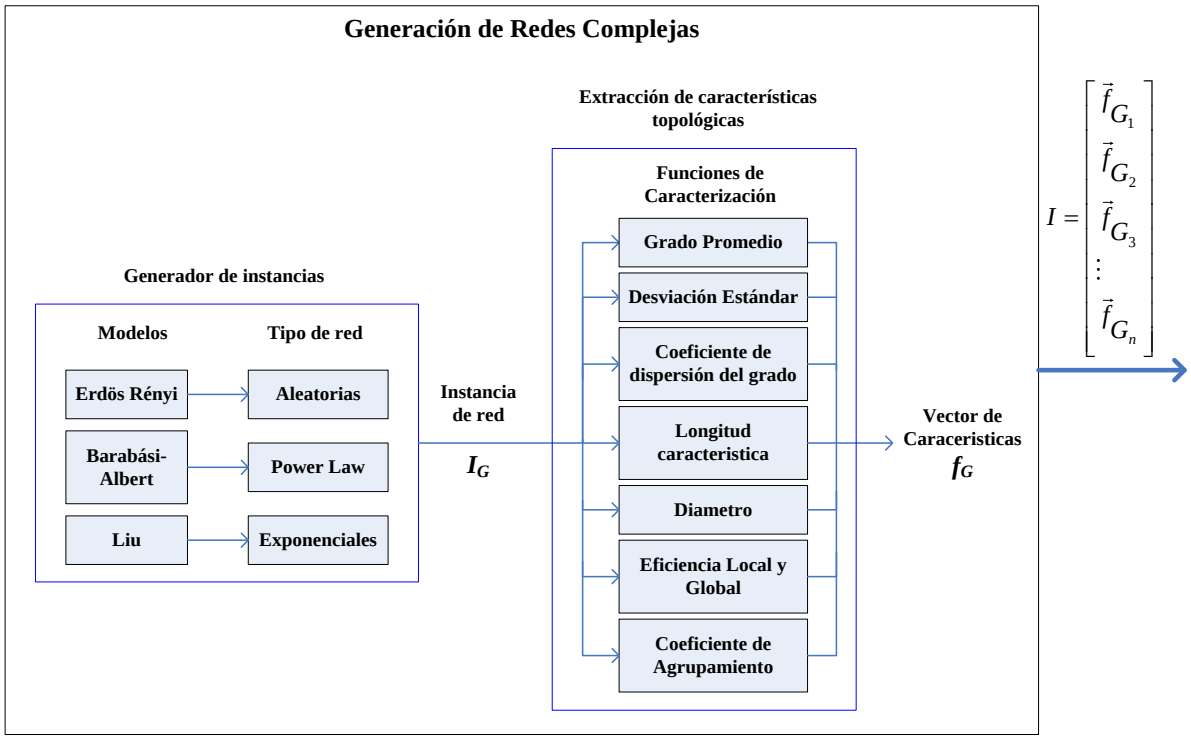


Figura 6.3 Generación de Redes Complejas

6.2.1 Generación de instancias de red.

En esta actividad se generaron instancias de redes Aleatorias, Power-Law y Exponenciales. Para generar redes Aleatorias se utilizó el modelo Erdös y Rényi, para generar redes Power-Law el modelo Barabási-Albert y para generar redes Exponenciales el modelo de Liu.

El número de nodos con el cual se generaron estos tipos de redes fue 200, 512, 1024, 2048 y 4096 nodos por cada tipo de red. En los trabajos relacionados los investigadores fijan el número de nodos y el número de aristas, de modo que el grado promedio es el mismo para todas las instancias de red que utilizan; en este trabajo no se fija el número de aristas con el propósito de tener diferente grado promedio en las instancias de red generadas. El número de instancias se determinó estadísticamente, este procedimiento se explica en el capítulo de experimentación.

En el caso de los modelos de Barabási y de Liu que requieren de parámetros que determinan el número de aristas en las redes como el número de nodos iniciales m_0 y el

número de aristas *m* que cada nuevo nodo establece, se determinó $m = m_0$, valores de 3, 5 y 7 se asignaron aleatoriamente a *m*, para tener diferente grado promedio en las redes.

6.2.2 Extracción de características topológicas.

Una vez generadas las instancias de red se extrajeron características topológicas globales mediante funciones de caracterización basadas en información global (ver sección 2.3), las funciones empleadas fueron: el grado promedio de la red $\langle k \rangle$, la desviación estándar del grado $\sigma_{\langle k \rangle}$, el coeficiente de dispersión del grado global *DDC*(*G*), la longitud de ruta característica *L*(*G*), el diámetro de la red *D*(*G*), la eficiencia local E_{loc}, la eficiencia global E_{glob}, y el coeficiente de agrupamiento *C*(*G*), que son funciones de caracterización [Lu 2005, Costa 2007] que reflejan la estructura topológica global de una red compleja

Al aplicar estas funciones sobre una instancia de red I_G se obtiene un vector de características $\vec{f}_G$ el cual se aprecian en la Figura 6.4: el primer campo contiene el nombre de la instancia del cual se extrajeron las características topológicas, los 8 siguientes campos contienen los valores de las funciones de caracterización mencionadas anteriormente, y el ultimo campo contiene una etiqueta que indica el tipo red, si es aleatoria la etiqueta tiene el valor de 1, si es Power-Law el valor de 2 y si es exponencial el valor de 3.

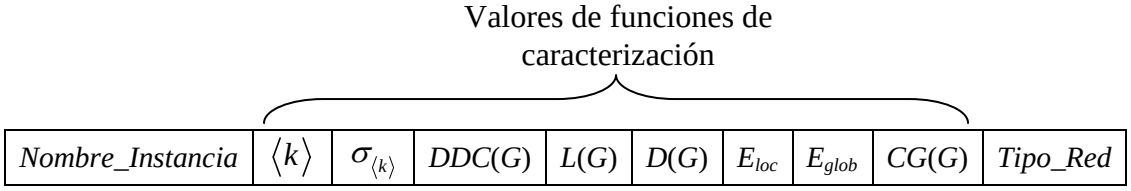


Figura 6.4 Estructura del vector de características

El objetivo de extraer características globales en esta etapa, es el de capturar el comportamiento global de las redes reflejado en estas características topológicas, debido que; si globalmente se puede observar que los valores de ciertas funciones de caracterización son diferentes dependiendo del tipo de red, creemos entonces que al aplicar

estas funciones a cada subgrafo de la red tomado con un todo, se podrá distinguir que tipo de red es localmente.

6.3 SELECCIÓN DE CARACTERÍSTICAS

Una vez que se generaron y caracterizaron las instancias de red, se procedió a seleccionar las funciones de caracterización que nos permitan discriminar entre redes Aleatorias, Power-Law y Exponenciales.

El objetivo de la selección es encontrar el conjunto de funciones de caracterización que mejor contribuyan a discriminar entre las redes Aleatorias, Power-Law y Exponenciales, dejando a un lado aquellas funciones que su contribución a la discriminación sea escasa, bien porque no son relevantes o porque resulten redundantes.

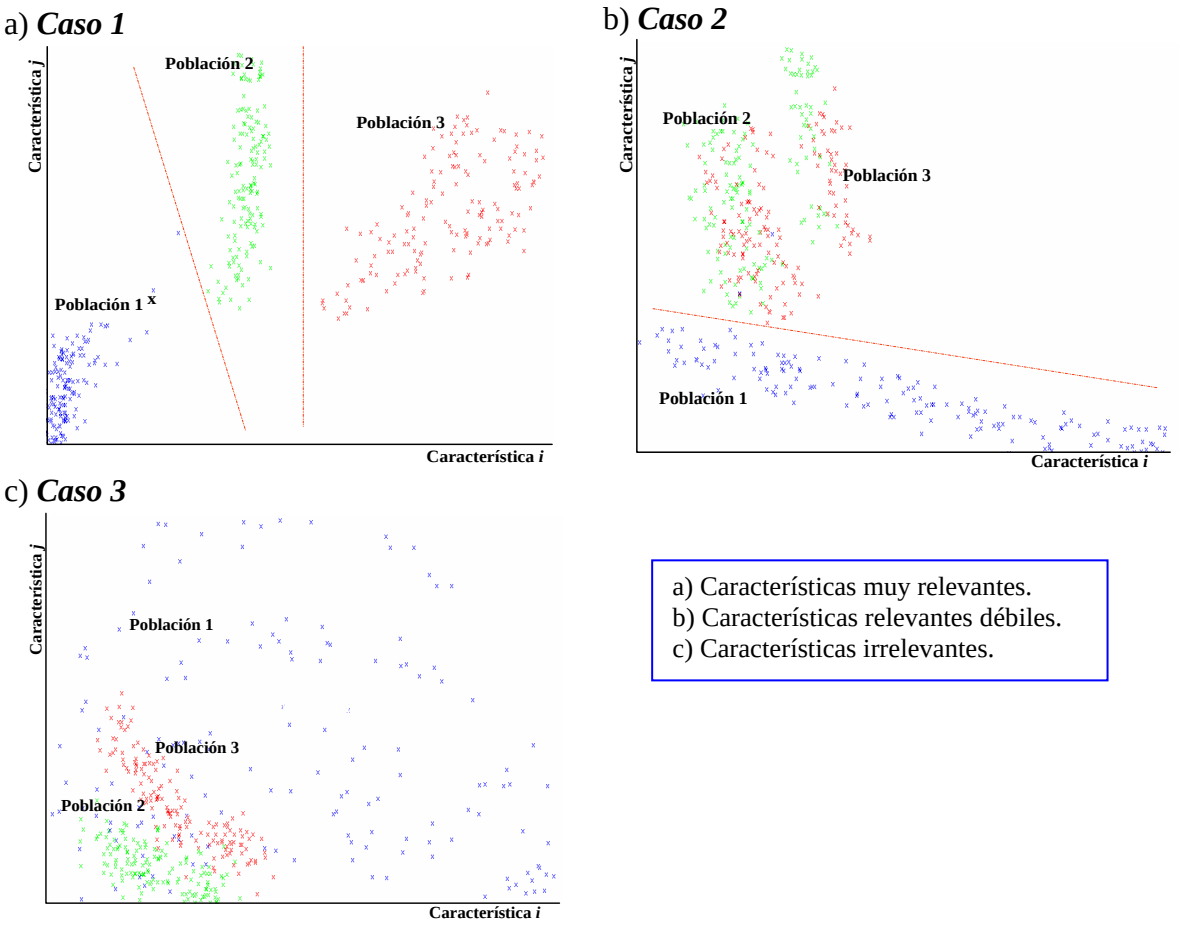


Figura 6.5 Diferencias significativas entre tipos de redes

Para lograr este objetivo se siguió un enfoque estadístico, en donde cada tipo de red se considera una población con diferentes características topológicas, si las medias de esas características son diferentes y las poblaciones no se traslapan, entonces esas características pueden ayudar a distinguir redes de esas poblaciones. De esto se pueden esperar tres posibles casos, ilustrados en la Figura 6.5 y los cuales se detallan a continuación:

- ***Caso 1***: Existen diferencias significativas entre los tres tipos de red respecto a un conjunto de características, es así que mediante esas características se puede clasificar una nueva red dentro de un tipo de red, Figura 6.5 a). Estas características son muy relevantes.
- ***Caso 2***: Existen diferencias significativas entre algunos de los tipos de red respecto a un conjunto de características, Figura 6.5 b), esas características pueden conducir a clasificaciones erróneas de nuevas redes. Estás son características relevantes débiles.
- ***Caso 3***: No existen diferencias significativas entre los tres tipos de red respecto a un conjunto de características, Figura 6.5 c), esas características conducen a clasificaciones erróneas de nuevas redes. Estas características son irrelevantes.

En esta etapa a partir de un conjunto de funciones de caracterización $F = \{\langle k \rangle, \sigma_{\langle k \rangle}, DDC(G), L(G), D(G), E_{loc}, E_{glob}, C(G)\}$, se obtiene el conjunto mínimo F' de funciones de caracterización que contiene las funciones muy relevantes y algunas relevantes débiles, sin que haya redundancia entre ellas.

Para ello se realizaron cuatro actividades: identificar características relevantes e irrelevantes, a partir de las características relevantes identificar las características muy relevantes y las relevantes débiles, eliminar la redundancia entre las características, e identificar el conjunto mínimo de funciones de caracterización. Estas actividades, se esquematizan en la Figura 6.6.

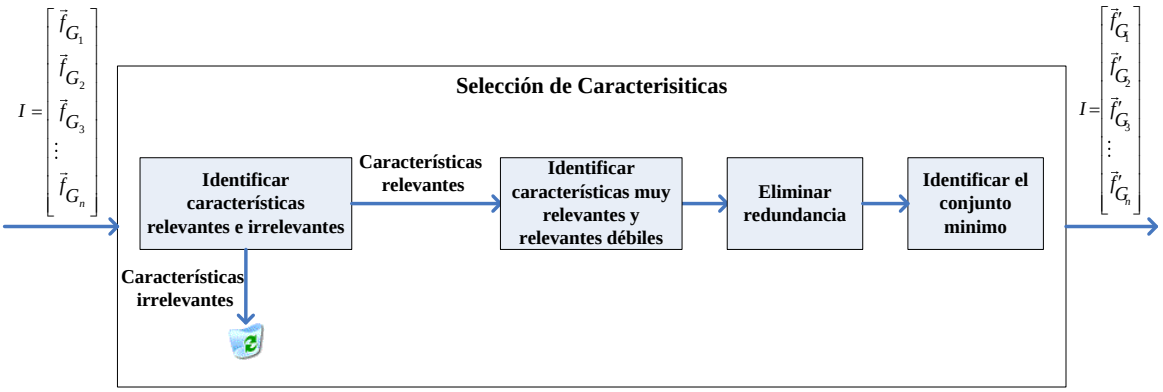


Figura 6.6 Selección de características

6.3.1 Identificación de características relevantes e irrelevantes.

Esta actividad tiene como objetivo identificar en qué funciones de caracterización del conjunto F las medias son diferentes de acuerdo al tipo de red, no importando el número de nodos que existan en la red.

Las funciones de caracterización cuyas medias sean diferentes son consideras como características relevantes (*Caso 1* y *2*) y las que no difieran son irrelevantes (*Caso 3*). Para identificar estos casos se diseñó un experimento siguiendo el esquema general de procedimientos sugerido en [Montgomery 2004].

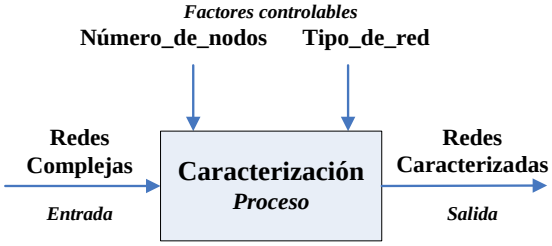


Figura 6.7 Modelo del proceso de caracterización de redes complejas.

Existen diferentes factores que influyen en el proceso de caracterización de las redes complejas como: el tipo de red, el número de nodos y el número de aristas existentes en la red. Para este experimento se seleccionaron dos factores que se pueden controlar sin afectar

los supuestos de normalidad, el tipo de red y el número de nodos en la red. En la Figura 6.7 se representa el modelo del proceso de caracterización de redes complejas.

Dadas las características del problema se eligió un diseño factorial con dos factores para detectar diferencias significativas de acuerdo al tipo de red en las funciones de caracterización mencionadas en la sección 6.1. En la experimentación se decidirá el modelo con el cual se trabajará en el diseño factorial.

6.3.2 Identificación de características muy relevantes y relevantes débiles

En esta actividad se toman las funciones de caracterización identificadas como relevantes y se realiza un análisis de las funciones relevantes mediante múltiples comparaciones entre ellas, para determinar cuales son características muy relevantes y características relevantes débiles.

Específicamente se desea detectar que funciones tienen medias que difieren para las redes Aleatorias, Power-Law y Exponenciales e identificar que funciones concuerdan con el Caso 1 o Caso 2 descritos anteriormente.

6.3.3 Eliminación de características redundantes

En esta actividad se eliminan aquellas funciones de caracterización que resulten redundantes, realizando un análisis de correlación.

El objetivo de esta actividad es eliminar en primera instancia las características relevantes débiles que estén muy relacionadas con las características muy relevantes, de esta manera el conjunto de funciones de caracterización seleccionadas contendrá las características muy relevantes y algunas relevantes débiles.

6.3.4 Identificación del conjunto óptimo

En esta actividad se realizan combinaciones de las características seleccionadas, haciendo uso del análisis discriminante se determinará la combinación con el menor número de

funciones de caracterización que produzca mejores resultados en el análisis discriminante, a esta combinación se le llamará conjunto óptimo F'.

6.4 CLASIFICACIÓN DE REDES COMPLEJAS

En esta etapa el objetivo es determinar el clasificador con mejor desempeño empleando como entradas al clasificador las funciones de caracterización del conjunto optimo F' y usar ese clasificador para clasificar nuevas instancias de red.

Esta etapa consta de tres actividades, como puede apreciarse en la Figura 6.8, la determinación del clasificador con mejor desempeño, clasificar instancias de red de naturaleza desconocida e identificar el tipo de red para cada subgrafo de instancias de Internet.

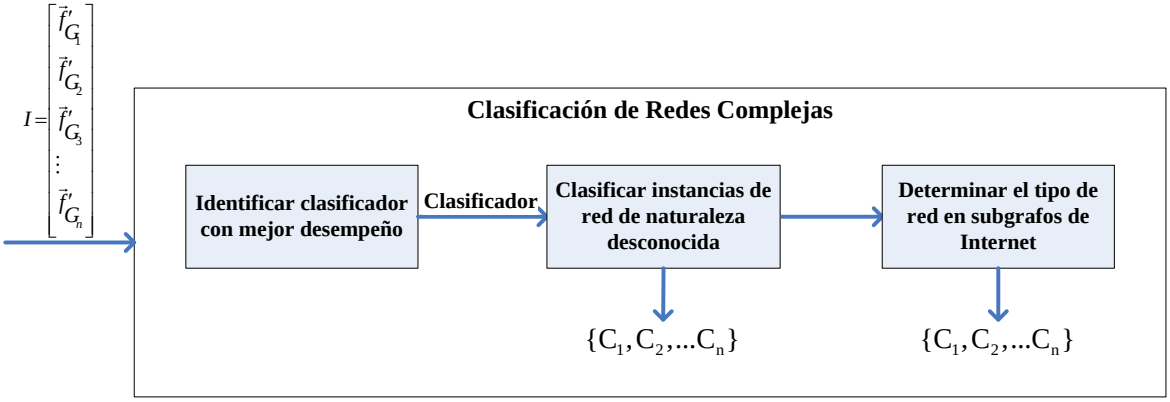


Figura 6.8 Clasificación de Redes Complejas

6.4.1 Identificación del clasificador con el mejor desempeño.

En esta actividad se utilizan las funciones de caracterización del conjunto mínimo F' y alguna herramienta utilizada en el área de aprendizaje automático, para identificar el clasificador con el mejor desempeño.

6.4.2 Clasificación de instancias de naturaleza desconocida

Para esta actividad se establece contacto con investigadores del área de redes complejas para solicitar instancias de red de las cuales no se conoce el tipo de red.

Estas instancias de red son caracterizadas con las funciones de caracterización del conjunto mínimo F' y se determina el tipo de red empleando el clasificador identificado en la actividad descrita en la sección 6.4.1, posteriormente se comprueban los resultados obtenidos en el proceso de clasificación, con la fuente que proporciono las instancias para calcular el porcentaje de aciertos obtenidos.

6.4.3 Determinación del tipo de red en cada nodo de las instancias de Internet

En esta actividad se toman instancias de Internet que están disponibles en Sitios Web de organizaciones se que tienen como objetivo el estudio de las redes complejas y de la Internet (ver Anexo A).

En estas instancias se caracteriza cada subgrafo asociado a cada nodo de la red con las funciones de caracterización del conjunto mínimo F' y se identifica el tipo de red utilizando el clasificador con el mejor desempeño.

Capítulo 7

EXPERIMENTACIÓN Y ANÁLISIS DE RESULTADOS

En el capitulo 6 se propone una metodología basada en un agente de aprendizaje, donde se plantea el uso de un enfoque estadístico y técnicas de aprendizaje supervisado. En este capítulo se describe la aplicación de la metodología descrita en el capítulo anterior que tiene como objetivo identificar un conjunto de funciones de caracterización que permitan discriminar entre diferentes topologías de redes complejas, como las redes Aleatorias, redes Power-Law y redes Exponenciales.

El capítulo está organizado de la siguiente manera: en la sección 7.1 se describe como se generaron y caracterizaron las redes complejas analizadas, los modelos y funciones de caracterización empleadas y como se determino la cantidad de redes a generar; en la sección 7.2 se muestra el análisis estadístico realizado a las funciones de caracterización y se identifica la función de caracterización que mejor discrimina entre redes Aleatorias, Power-Law y Exponenciales; en la sección 7.3 se describe como se determino el clasificador con mejor desempeño tomando como entrada la función de caracterización identificada en la sección 7.2, se determina el tipo de red de instancias de redes de las cuales se desconoce su tipo y se realiza la identificación del tipo de red en cada nodo de instancias de red.

7.1 GENERACIÓN DE REDES COMPLEJAS

El generador de redes complejas fue implementado en una primera versión en MATLAB 7 a como un prototipo para analizar la factibilidad del proyecto; la versión definitiva se implementó en Visual Studio C++ 6.0, siendo con esta versión en la que se generaron y caracterizaron las instancias de red.

Las características del equipo en el cual se generaron y caracterizaron las instancias de red fue en un servidor con sistema operativo Windows Server 2003, 2 procesadores Xenon (TM) CPU 3.06GHz con 3.87GB de memoria RAM.

Una decisión crítica en cualquier experimentación es la elección del tamaño de la muestra, es decir, el número de réplicas que deben efectuarse del experimento. Si se tiene interés en detectar diferencias pequeñas entre las medias de las variables a medir, se necesitan más replicas que cuando se interesa detectar diferencias grandes.

En este caso el tamaño de la muestra es el número de instancias a generar, la generación de redes complejas y su caracterización implica mucho tiempo computacional por lo que determinar el número de instancias a generar es una decisión importante en esta etapa del proyecto.

Para determinar el número de instancias es necesario determinar los niveles de los factores tipo de red y número de nodos detectados en la sección 6.3.1. El factor *A* tipo de red, consta de tres niveles *a*, representados por los tres tipos de redes que se están analizando, redes Aleatorias (1), Power-Law (2) y Exponenciales (3). El factor *B* número de nodos, consta de cinco niveles *b*, 200, 512, 1024, 2048 y 4096 nodos.

Dadas las características del problema, en la sección 6.3.1 se decidió elegir un diseño factorial con dos factores, el procedimiento adecuado para determinar el tamaño de la muestra en este tipo de diseño experimental se puede encontrar en [Montgomery 2004], a continuación se presentan los cálculos realizados para determinar el número de instancias de red a generar.

7.1.1 Determinación del número de instancias de red.

Para determinar el número de instancias de red, apropiadas para detectar diferencias significativas entre las medias de los niveles de los factores, se hizo uso de las curvas de operación característica. El emplear estas curvas de operación características consiste encontrar el menor valor para el parámetro Φ^2 en el modelo de los efectos fijos y para el parámetro λ en el modelo mixto, que tienen relación con las diferencias que se desean detectar entre las medias de los niveles de los factores [Montgomery 2004].

Tabla 7.1 Fórmulas para determinar del tamaño de la muestra

Diseño factorial con dos factores, modelo de los efectos fijos.			
Factor	Φ^2	Grados de libertad del numerador	Grados de libertad del denominador
A	$\Phi^2 = \frac{nbD^2}{2a\sigma^2}$	$a-1$	$ab(n-1)$
B	$\Phi^2 = \frac{naD^2}{2b\sigma^2}$	$b-1$	$ab(n-1)$
AB	$\Phi^2 = \frac{nD^2}{2\sigma^2[(a-1)(b-1)+1]}$	$(a-1)(b-1)$	$ab(n-1)$

Diseño factorial con dos factores, modelo mixto.			
Factor	Parámetro	Grados de libertad del numerador	Grados de libertad del denominador
A	$\Phi^2 = \frac{nbD^2}{2a[\sigma^2 + n\sigma^2_{\tau\beta}]}$	$a-1$	$(a-1)(b-1)$
B	$\lambda = \sqrt{1+\frac{an\sigma^2_{\beta}}{\sigma^2}}$	$b-1$	$ab(n-1)$
AB	$\lambda = \sqrt{1+\frac{n\sigma^2_{\tau\beta}}{\sigma^2}}$	$(a-1)(b-1)$	$ab(n-1)$

Los valores mínimos de los parámetros Φ^2 y λ para los factores se calculan mediante las fórmulas de la Tabla 7.1. Las variables en las fórmulas significan: *n* = el número de réplicas, *a* = el número de niveles del factor *A*, *b* = el número de niveles del factor *B*, *D* = la diferencia que se desea detectar y los parámetros σ^2 corresponden a la varianza entre los niveles de los factores, para una mejor comprensión de estas fórmulas y como emplearlas junto con las curvas de operación características, véase [Montgomery 2004].

Estas fórmulas también permiten identificar la potencia del experimento es decir, la probabilidad de rechazar la hipótesis nula si no se detecta la diferencia indicada. Para corroborar los cálculos y graficar la potencia del experimento versus el número de instancias se utilizó la herramienta Java Applets for power and sample size [Lenth 2006].

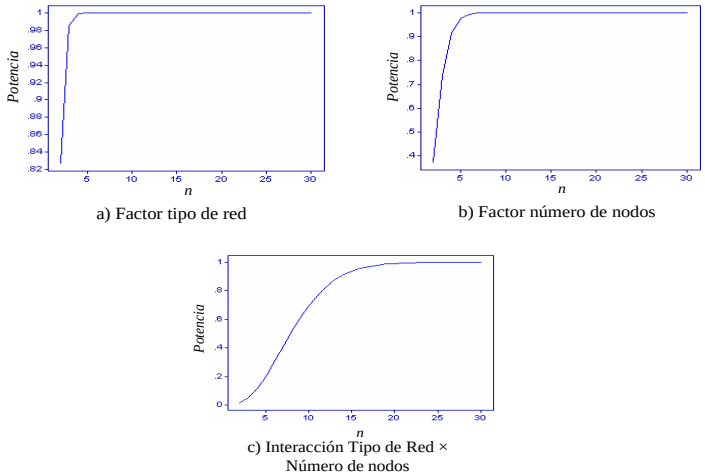
Los resultados para el modelo de efectos fijos, se muestran en la Tabla 7.2 que contiene los mejores valores de *n* para maximizar la potencia del experimento en cada uno de los factores y la interacción. El parámetro *D* se fijo en 0.1 debido que se desean detectar diferencias muy pequeñas entre las medias de los niveles de los factores. El nivel de significación α (probabilidad de que ocurra el error tipo I) se fijó en 0.01, con los parámetros descritos anteriormente se obtuvo el valor de β (probabilidad de que ocurra el error tipo II) y la potencia (1-β) del experimento para detectar diferencias pequeñas entre las medias.

Tabla 7.2 Determinación del tamaño de la muestra para el modelo de efectos fijos

	Parámetros						Resultados	
Factor	*a*	*b*	σ^2	*D*	***n***	α	β	(1-β)
Tipo de red	3	5	0.05	0.1	**3**	0.01	0.0147	0.9853
Número de nodos	3	5	0.05	0.1	**6**	0.01	0.0055	0.9945
Interacción	3	5	0.05	0.1	**19**	0.01	0.0128	0.9872

Mediante este procedimiento se puede observar que el número de réplicas *n* se puede fijar en 19 obteniéndose un riesgo β menor al 1.5% de aceptar la hipótesis nula dado que es falsa, la potencia del experimento tendrá por lo menos un 98% de probabilidad de detectar diferencias de hasta 0.1 entre las medias de los niveles de los factores y la interacción. Se puede concluir, que si se usan 19 instancias de red por cada tipo de red con diferente números de nodos bastan para proporcionar la sensibilidad deseada al experimento, tomando en cuenta que los efectos de los factores son fijos.

Debido a que el experimento fue extendido a la clasificación de instancias $n=19$ fue considerado como el número mínimo de instancias a generar para cada tipo de red con diferente número de nodos; para la selección de características se fijo $n=30$. La potencia del experimento cuando $n=30$ es muy cercana al 100%, esto se puede observar en la Figura 7.1, en donde se muestra como después de la *n* calculada para los factores y la interacción mostradas en la Tabla 7.2 la potencia tiende a ser constante para cualquier valor mayor que la *n* calculada.



a) Factor tipo de red

b) Factor número de nodos

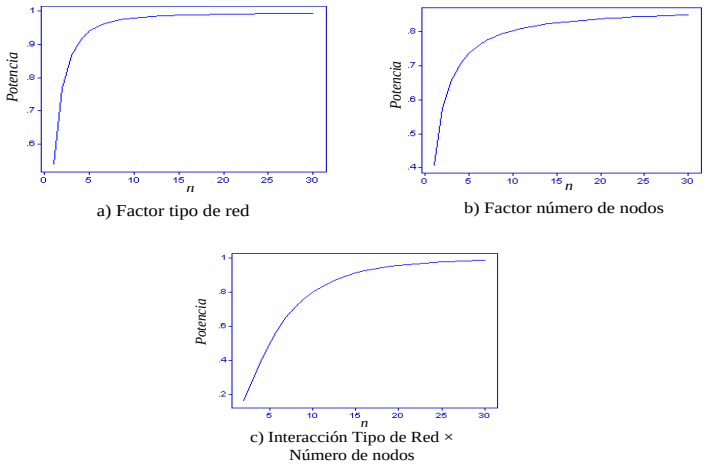
c) Interacción Tipo de Red × Número de nodos

Figura 7.1 Gráficas Potencia vs Número de instancias. Modelo de efectos fijos.

Se fijó *n*=30 y se calculó la potencia del experimento cuando el modelo se considera mixto, los resultados se muestran en la Figura 7.2, debido a que se considera uno de los factores como aleatorio, el parámetro D se fijó en 0.2.

Tabla 7.3 Determinación del tamaño de la muestra para el modelo mixto

	Parámetros								Resultados	
Factor	A	b	σ^2	σ^2_β	$\sigma^2_{\tau\beta}$	D	$\boldsymbol{n}$	α	β	$(1-\beta)$
Tipo de red	3	5	0.1			0.2	**30**	0.05	0.0064	0.9936
Número de nodos	3	5	0.1	0.1		0.2	**30**	0.05	0.1504	0.8496
Interacción	3	5	0.1		0.05	0.2	**30**	0.05	0.0148	0.9852



a) Factor tipo de red

b) Factor número de nodos

c) Interacción Tipo de Red × Número de nodos

Figura 7.2 Gráficas Potencia vs Número de instancias. Modelo mixto.

Se puede observar que fijando el número de réplicas *n* en 30 se obtiene un riesgo β menor al 15% de aceptar la hipótesis nula dado de que es falsa, la potencia del experimento tendrá por lo menos un 84% de probabilidad de detectar diferencias de hasta 0.2 entre las medias de los niveles de los factores y las interacciones en el modelo mixto, en la Figura 7.2 se muestra el comportamiento de la potencia del experimento, comparándose con las gráficas de la Figura 7.1 se puede observar que se necesitan un número mayor de instancias para detectar diferencias pequeñas en las medias en un modelo mixto.

7.2 SELECCIÓN DE CARACTERÍSTICAS

Es necesario recordar que el objetivo de la selección de características es encontrar el conjunto de funciones de caracterización que mejor contribuyan a discriminar entre las redes Aleatorias, Power-Law y Exponenciales.

Se realizaron análisis gráficos de la distribución de los valores de las funciones de caracterización utilizando el software estadístico Minitab 14. El objetivo de este análisis es visualizar que efecto (de localización y/o dispersión) tienen los factores tipo de red y número de nodos sobre las funciones de caracterización. Enfocaremos nuestra atención en funciones de caracterización en las cuales los niveles del factor tipo de red afecten la localización de la media y/o la dispersión de las funciones de caracterización, de manera que la media sea diferente y los datos no se traslapen y que los niveles del factor número de nodos no afecte la localización de media y dispersión de los datos.

En la Figura 7.3 se muestra el histograma de la distribución de los valores de la función *DDC*(*G*) ajustados a la distribución normal según el tipo de red no importando el número de nodos presente en la red, donde se puede apreciar que esta función de caracterización difiere de acuerdo al tipo de red.

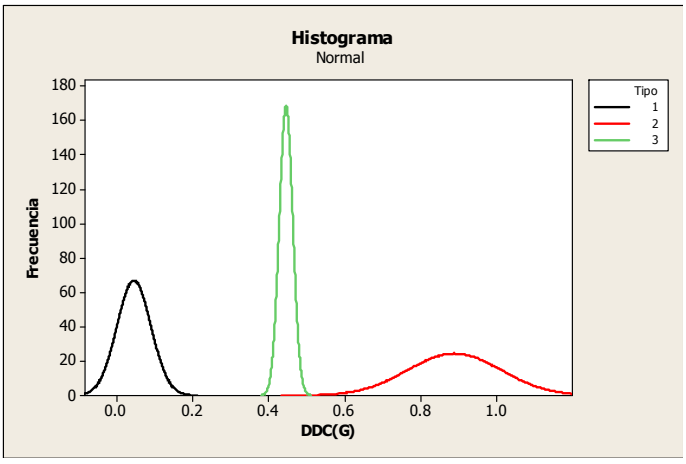


Figura 7.3 Histograma de la función *DDC*(*G*) por tipo de red.

En la Tabla 7.4 se muestra la media, desviación estándar, el valor mínimo y el valor máximo de la función *DDC*(*G*), se puede observar que las medias son diferentes de acuerdo al tipo red, la desviación estándar es pequeña por lo que los datos no están muy alejados de la media, y los valores de máximos y mínimos dan indicios que la distribución de los valores del *DDC*(*G*) de acuerdo al tipo de red no se traslapan. Se puede decir que para esta función el factor tipo de red afecta tanto la localización de la media como la dispersión.

Tabla 7.4 Estadísticos descriptivos de la función *DDC*(*G*) por tipo de red.

DDC(*G*)				
Tipo de Red	μ	σ	**Min**	**Max**
1	0.0458	0.0447	0.0015	0.3025
2	0.8888	0.1227	0.6121	1.1118
3	0.4458	0.0177	0.3904	0.4718

Por otra parte, en la Figura 7.4 se muestra el histograma de la distribución de los valores del *DDC*(*G*) ajustados a la distribución normal según el número de nodos no importando el tipo de red, donde se puede apreciar que esta función de caracterización no difiere de acuerdo al número de nodos.

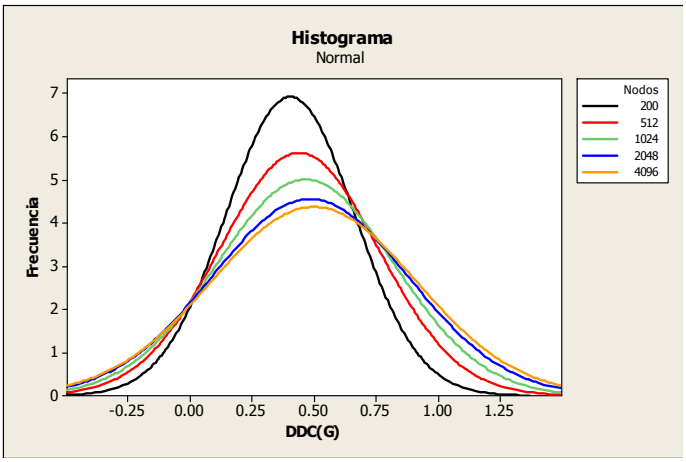


Figura 7.4 Histograma de la función *DDC*(*G*) por número de nodos.

En la Tabla 7.5 se muestran la media, desviación estándar, el valor mínimo y el valor máximo de esta función, se puede observar que las medias tienen valores similares, la desviación estándar es grande, lo que significa que los datos están muy alejados de la media, y los valores de máximos y mínimos dan indicios que la distribución de los valores del *DDC*(*G*) de acuerdo al número de nodos se traslapan. Se puede decir que para esta función el factor tipo de número de nodos no afecta la localización de la media ni la dispersión.

Tabla 7.5 Estadísticos descriptivos de la función *DDC*(*G*) por número de nodos.

***DDC*(*G*)**				
Número de nodos	μ	σ	**Min**	**Max**
200	0.4057	0.2595	0.0133	0.8235
512	0.4413	0.3194	0.00766	0.9584
1024	0.4670	0.3588	0.00561	0.9718
2048	0.4858	0.3945	0.00247	1.0616
4096	0.5011	0.4111	0.00151	1.1118

La distribución de los valores de las funciones de caracterización restantes se encuentra en el Anexo B; en el Cuadro b1 del Anexo B se muestran los estadísticos descriptivos para las funciones de caracterización según el tipo de red, y en el Cuadro b2 del Anexo B se muestran los estadísticos descriptivos para las funciones de caracterización según el número de nodos, así también en este anexo se encuentran los histogramas de las distribuciones de las funciones de caracterización por tipo de red y por número de nodos.

Observando y analizando esta información se puede suponer que la función *DDC*(*G*) podría ayudar a discriminar entre los tipos de red independientemente de la cantidad de nodos, debido a que la media según el tipo de red se ve afectada en su localización y la dispersión de los valores para diferentes tipos de red no se traslapan, esta suposición se podría extender para cualquier cantidad de nodos debido a que este factor no afecta la media ni la dispersión de esta función.

Para corroborar este supuesto se llevo acabo la ejecución del diseño factorial con dos factores y de esta manera identificar qué funciones de caracterización son relevantes o irrelevantes.

7.2.1 Identificación de características relevantes e irrelevantes.

En este paso es de interés identificar en qué funciones de caracterización, la media sea significativamente diferente de acuerdo al tipo de red, no importando el número de nodos

que existan en la red. Para ello se efectuaron dos experimentos basados en el diseño factorial con dos factores, utilizando el software estadístico Minitab 14.

En términos de nuestro problema, si las medias de las funciones de caracterización son significativamente diferentes de acuerdo al tipo de red, no importando el número de nodos en la red, esas funciones de caracterización pueden contribuir a discriminar entre los tipos de redes aquí tratados. Las funciones que cumplan con esta condición serán catalogadas como características relevantes y las que no cumplan esta condición serán catalogadas como características no relevantes.

En el primero de los experimentos se consideró que los efectos de los factores son fijos lo que lleva a conclusiones válidas sólo para los niveles de los factores que se están considerando, en este experimento se trabajo con el modelo de los efectos fijos; en el segundo experimento se considera que el factor número de nodos es aleatorio con el objetivo de extender las conclusiones para cualquier cantidad de nodos, el modelo utilizado es el mixto. En este punto es necesario recordar los datos con los cuales se lleva a cabo las experimentaciones:

Factores:

A = Tipo de red — B = Número de nodos

Niveles $a = 3$ — Niveles $b = 5$

Variables respuestas (funciones de caracterización): $p = 8$

$\langle k \rangle$ = Grado promedio de la red.

$\sigma_{\langle k \rangle}$ = Desviación estándar del grado.

$DDC(G)$ = Coeficiente de dispersión del grado.

$L(G)$ = Longitud de ruta características más corta.

$D(G)$ = Diámetro de la red.

E_{loc} = Eficiencia local de la red.

E_{glob} = Eficiencia global de la red.

$CG(G)$ = Coeficiente de agrupamiento global.

7.2.1.1 Diseño factorial modelo de los efectos fijos

De acuerdo al modelo de los efectos fijos que describe el diseño factorial con dos factores fueron formuladas hipótesis para detectar diferencias significativas para cada una de las funciones de caracterización.

Las hipótesis nulas se plantean de manera que, si no existen diferencias significativas entre las medias de los niveles de los factores, los efectos de esos niveles del factor son cero; de lo contrario las hipótesis alternativas planean que, si existen diferencias significativas al menos un efecto de los niveles en los factores debe ser diferente de cero.

En la Tabla 7.6 se muestran las hipótesis planteadas para los efectos principales y la interacción con su correspondiente criterio de rechazo y los valores de tablas con los cuales se rechazarán o aceptarán las hipótesis. El nivel de significación α se fijo en 0.01.

Tabla 7.6 Hipótesis asociadas al modelo de los efectos fijos

Hipótesis	Criterio de rechazo	Valor de Tablas
$H_0 : \tau_1 = \tau_2 = \cdots \tau_a = 0$ H_1 : al menos una $\tau_i \neq 0$	$F_0 > F_{\alpha,a-1,(N-1)ab}$	$F_{0.01,2,435} = 4.61$
$H_0 : \beta_1 = \beta_2 = \cdots \beta_b = 0$ H_1 : al menos una $\beta_i \neq 0$	$F_0 > F_{\alpha,b-1,(N-1)ab}$	$F_{0.01,4,435} = 3.32$
$H_0 : (\tau\beta)_{ijl} = 0$ para todas las i, j H_1 : al menos una $(\tau\beta)_{ijl} \neq 0$	$F_0 > F_{\alpha,(a-1)(b-1),(N-1)ab}$	$F_{0.01,8,435} = 2.51$

Un Análisis Multivariado de la Varianza General (GLM ó MANOVA) fue llevado acabo para obtener resultados que permiten rechazar o aceptar las hipótesis nulas asociadas al modelo de los efectos fijos, detalles de los resultados se encuentran en el Cuadro c1 del Anexo C, los valores F_0 calculados se presentan en la

Tabla 7.7 y se discuten a continuación.

Tabla 7.7 Valores F_0 calculados mediante GLM para cada función de caracterización.

Factor	Funciones de caracterización							
	$\langle k \rangle$	$\sigma_{\langle k \rangle}$	$DDC(G)$	$L(G)$	$D(G)$	E_{loc}	E_{glob}	$CG(G)$
Tipo de red	218.83	327.73	18794.1	1348.78	1014.56	95.73	1066.68	406.02
Núm. de nodos	39.02	93.33	90.40	77.22	32.18	**1.87**	15.59	**2.41**
Interacción	38.93	46.71	128.51	22.48	11.90	**0.54**	4.35	**0.94**

Los valores mostrados en la Tabla 7.7, fueron comparados con los criterios de rechazo asociados con las hipótesis formuladas, se puede observar que de acuerdo al tipo de red las funciones de caracterización cumplen con el criterio de rechazo, es decir, las hipótesis nulas se rechazan, por tanto, se concluye que todas las funciones de caracterización difieren significativamente de acuerdo al tipo de red.

De acuerdo al número de nodos y la interacción entre el tipo de red y el número de nodos, las funciones de caracterización que no cumplen con el criterio de rechazo son E_{loc} y $CG(G)$, por lo que se concluye que las medias de la eficiencia local y el coeficiente de agrupamiento son significativamente iguales en este factor y en la interacción.

Es importante poner atención a los valores F_0 para el factor tipo de red, al ser estos mayores que los valores F_0 del factor número de nodos y de la interacción, se concluye que el factor tipo de red es el que tiene mayor influencia en los datos obtenidos con las funciones de caracterización, sobre todo para las funciones $DDC(G)$, $L(G)$, $D(G)$ y E_{glob}.

Tabla 7.8 Valores F_0 calculados por la prueba MANOVA y valores de la distribución F.

Prueba MANOVA	Tipo de red		Número de nodos		Interacción	
	F_0	F_{α,ν_1,ν_2}	F_0	F_{α,ν_1,ν_2}	F_0	F_{α,ν_1,ν_2}
Wilks'	3355.989	2.04	73.621	1.7	48.63	1.47
Pillilla's	2826.861	2.04	28.422	1.7	20.370	1.47

Lawley – Hotelling	3982.246	2.04	219.904	1.7	103.89	1.47

Los valores aproximados de F_0 obtenidas mediante las pruebas MANOVA mostradas en la Tabla 7.8, para el tipo de red, el número de nodos y la interacción, son estadísticamente significativas, por lo que se reafirma que las diferencias significativas que se detectan en los análisis de una variable a la vez de la

Tabla 7.7 son reales y no falsas positivas. Debido a que los valores de F_0 son grandes para el tipo de red, se concluye nuevamente que este factor tiene mayor influencia en los valores obtenidos por las funciones de caracterización.

7.2.1.2 Diseño factorial modelo mixto

De acuerdo al modelo mixto que describe el diseño factorial con dos factores fueron formuladas hipótesis para detectar diferencias significativas para cada una de las funciones de caracterización.

Las hipótesis nulas del modelo mixto para el factor aleatorio y la interacción se plantean de manera diferente que las hipótesis nulas del modelo de los efectos fijos. En el caso del modelo mixto es de interés detectar si existe o no variabilidad en los niveles del factor aleatorio, por lo que las hipótesis para el factor aleatorio y la interacción se plantea en función de los componentes de la varianza, para una mejor comprensión del modelo mixto véase [Montgomery 2004].

En la Tabla 7.9 se muestran las hipótesis planteadas para los efectos principales y la interacción según el modelo mixto, con su correspondiente criterio de rechazo y los valores de tablas con los cuales se rechazaran o aceptarán las hipótesis. El nivel de significación α se fijo en 0.05.

Tabla 7.9 Hipótesis asociadas al modelo mixto

Hipótesis	Criterio de rechazo	Valor de Tablas
$H_0 : \tau_1 = \tau_2 = \cdots \tau_a = 0$ H_1 : al menos una $\tau_i \neq 0$	$F_0 > F_{\alpha,a-1,(a-1)(b-1)}$	$F_{0.05,2,435} = 4.46$

Continuación Tabla 7.10 Hipótesis asociadas al modelo mixto

$H_0 : \hat{\sigma}_{\beta}^2 = 0$ $H_1 : \hat{\sigma}_{\beta}^2 \neq 0$	$F_0 > F_{\alpha,b-1,(n-1)ab}$	$F_{0.05,4,435} = 2.37$
$H_0 : \hat{\sigma}_{\beta}^2 = 0$ $H_1 : \hat{\sigma}_{\tau\beta}^2 \neq 0$	$F_0 > F_{\alpha,(a-1)(b-1),(n-1)ab}$	$F_{0.05,8,435} = 1.94$

Un Análisis Multivariado de la Varianza Balanceado (AOV) fue llevado acabo para obtener resultados que permiten rechazar o aceptar las hipótesis nulas asociadas al modelo mixto, detalles de los resultados se encuentran en el Cuadro c2 del Anexo C, los valores F_0 calculados se presentan en la Tabla 7.11 y se discuten a continuación.

Tabla 7.11 Valores F_0 calculados mediante GLM para cada función de caracterización.

Factor	**Funciones de caracterización**							
	$\langle k \rangle$	$\sigma_{\langle k \rangle}$	$DDC(G)$	$L(G)$	$D(G)$	E_{loc}	E_{glob}	$CG(G)$
Tipo de red	5.62	7.02	146.24	60	85.22	175.77	244.98	430.99
Número de nodos	**1.0**	**2.0**	**0.70**	3.44	2.70	3.44	3.58	2.55
Interacción	38.93	46.71	128.51	22.48	11.90	**0.54**	4.35	**0.94**

Los valores mostrados en la Tabla 7.11, fueron comparados con los criterios de rechazo asociados con las hipótesis formuladas, se puede observar que de acuerdo al tipo de red las funciones de caracterización cumplen con el criterio de rechazo, es decir, las hipótesis nulas se rechazan y por tanto se concluye que todas las funciones de caracterización difieren significativamente de acuerdo al tipo de red.

De acuerdo al número de nodos las funciones de caracterización que cumplen con el criterio de rechazo son $L(G)$, $D(G)$, E_{loc}, E_{glob}, $CG(G)$, con lo cual se concluye que las medias de estas funciones de caracterización difieren significativamente de acuerdo al número de nodos en la red, por otro lado, las medias de las funciones de caracterización

$\langle k \rangle$, $\sigma_{\langle k \rangle}$ y $DDC(G)$ no difieren significativamente de acuerdo al número de nodos existentes en la red, es decir el número de nodos no afecta los valores obtenidos con las funciones $\langle k \rangle$, $\sigma_{\langle k \rangle}$ y $DDC(G)$.

Para la interacción entre los factores tipo de red y número de nodos, las funciones de caracterización que no cumplen con el criterio de rechazo son E_{loc} y $CG(G)$, por lo que se concluye que las medias de estas funciones son significativamente iguales en la interacción.

Un aspecto importante cuando se tiene un factor aleatorio es observar los componentes de la varianza, es decir, que tanta variabilidad introduce el factor aleatorio y la interacción a las variables respuestas, esto se analiza a continuación.

Tabla 7.12 Componentes de la varianza para cada función de caracterización.

Factor	**Funciones de caracterización**							
	$\langle k \rangle$	$\sigma_{\langle k \rangle}$	$DDC(G)$	$L(G)$	$D(G)$	E_{loc}	E_{glob}	$CG(G)$
Número de nodos	134	4.898	**-0.00060**	0.06986	0.1050	0.00010	0.00098	0.00043
Interacción	164628	14.406	0.00603	0.08223	0.1693	**-0.00011**	0.00087	**-0.00005**
Error	130209	9.455	0.00142	0.11486	0.4858	0.00697	0.00782	0.02644

Los componentes de la varianza de las funciones de caracterización para el factor número de nodos y la interacción se muestran en la Tabla 7.12, los valores negativos se asumen como ceros, debido a que los valores de P son grandes (ver Cuadro c2 del Anexo C) para estas estimaciones, se puede concluir que el factor número de nodos no introduce variabilidad a la función $DDC(G)$, y que la interacción entre el tipo de red y el número de nodos no introduce variabilidad para E_{loc} y $CG(G)$.

Por otro lado se observa que las funciones de caracterización con variabilidad grande son $\langle k \rangle$ y $\sigma_{\langle k \rangle}$, con esto se concluye que el número de nodos afecta considerablemente a estas funciones.

Tabla 7.13 Valores F_0 calculados por la prueba MANOVA y valores de la distribución *F*.

Prueba MANOVA	Tipo de red		Número de nodos		Interacción	
	F_0	F_{α,ν_1,ν_2}	F_0	F_{α,ν_1,ν_2}	F_0	F_{α,ν_1,ν_2}
Wilks'	180.977	2.04	5.947	1.7	48.63	1.47
Pillilla's	186.086	2.04	2.898	1.7	20.370	1.47
Lawley – Hotelling	0.000	2.04	-8.815	1.7	103.89	1.47

Los valores aproximados de F_0 obtenidas mediante las pruebas MANOVA mostradas en la Tabla 6.9, para el tipo de red, el número de nodos y la interacción, son estadísticamente significativas, además los valores de *P* son menores que el nivel de significación (ver Cuadro c2 del Anexo C), por lo que se reafirma que las diferencias significativas que se detectan en los análisis de una variable a la vez en la Tabla 7.11 son reales y no falsas positivas.

7.2.1.3 Análisis de las gráficas de los efectos.

Los resultados obtenidos por el análisis multivariado de la varianza pueden ser reforzados con el análisis de las gráficas de los efectos principales y de la interacción para cada función de caracterización. Mediante estás gráficas se puede observar el factor que tiene mayor efecto en las funciones de caracterización y que tanta interacción hay entre los factores.

En este problema en particular se buscan funciones de caracterización en las cuales, el factor tipo de red tenga un mayor efecto que el factor número de nodos, que no exista interacción entre los niveles del factor tipo de red en cada nivel del factor número de nodos, pero que si exista interacción entre los niveles factor número de nodos en cada nivel del factor tipo de red, es decir, que la función de caracterización refleje el comportamiento del tipo de red no importando el número de nodos de la red.

En la Figura 7.5, se muestra la gráfica de los efectos principales para la función *DDC*(*G*), en esta gráfica se puede observar que el factor tipo de red tiene mucho mayor efecto que el factor número de nodos, a medida que aumenta el número de nodos el *DDC*(*G*) se ve afectado ligeramente.

En la Figura 7.6, se muestra la gráfica de la interacción de los factores para la función *DDC*(*G*), se puede observar que la interacción entre el factor tipo de red y el factor número de nodos es débil, además se observa que en las redes Aleatorias (1) conforme aumenta el número de nodos la media de la función *DDC*(*G*) disminuye, en las redes Exponenciales (3) la media tiende a ser constante y en las redes Power-Law (2) la media aumenta. Las redes con diferente número de nodos interactúan dependiendo el tipo de red.

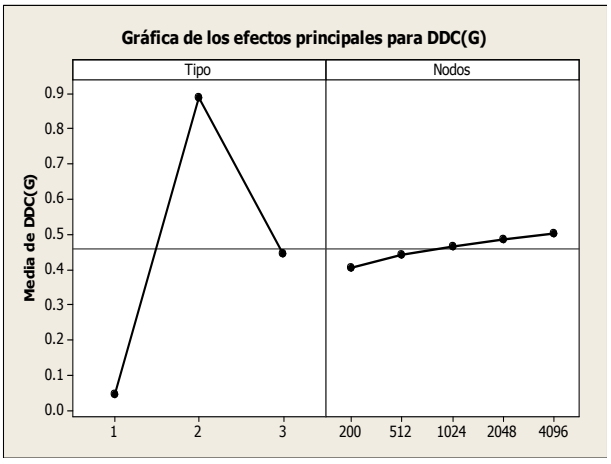


Figura 7.5 Gráfica de los efectos principales para la función *DDC*(*G*).

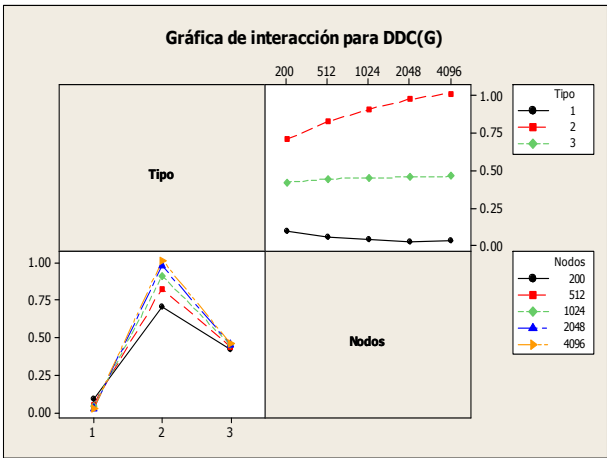


Figura 7.6 Gráfica de la interacción de factores para la función *DDC*(*G*).

Las gráficas de los efectos principales y la interacción de los factores para las funciones de caracterización restantes se encuentran en el Anexo D, en estas gráficas se puede observar que el tipo de red es el que tiene mayor efecto en todas las funciones de caracterización. El factor número de nodos tiene efectos grandes en funciones como $\langle k \rangle$, $\sigma_{\langle k \rangle}$, $L(G)$, $D(G)$ y efectos pequeños en funciones como E_{loc}, E_{glob}, $CG(G)$.

Con los resultados de las pruebas MANOVA para el modelo de los efectos fijos y el modelo mixto y con el análisis de las gráficas de los efectos e interacción de los factores, se puede concluir que la función de caracterización con mejores características para discriminar cuantitativamente entre las redes Aleatorias, Power-Law y Exponenciales es la función de caracterización $DDC(G)$.

7.2.1.4 Análisis de las gráficas de los residuales.

Para comprobar que el procedimiento del análisis de la varianza efectuado, es una prueba exacta de las hipótesis planteadas se debe verificar si el modelo de los efectos fijos y el modelo mixto describen adecuadamente las observaciones de las funciones de caracterización y que se satisfacen los supuestos de normalidad e independencia.

Las violaciones a estos supuestos y la adecuación del modelo pueden observarse a través del análisis de los residuales, por lo cual antes de adoptar las conclusiones realizadas

mediante el análisis multivariado de la varianza para el modelo de los efectos fijos y el modelo mixto, y el análisis de las gráficas de los efectos e interacción, se realizó un análisis de las gráficas de los residuales de las funciones de caracterización.

En la Figura 7.7 se observa la gráfica de los residuales para la función *DDC*(*G*), donde se puede apreciar que los residuales se distribuyen normalmente y no se observa ningún patrón obvio en los datos por lo que se puede decir que esta función satisface los supuestos de normalidad e independencia, de esta manera las conclusiones realizadas en la sección 7.2.1.3 son correctas.

En la Figura 7.8a) se muestra la gráfica de los residuales para la función *DDC*(*G*) por tipo de red y en la Figura 7.8b) se muestra la gráfica de los residuales para la función *DDC*(*G*) por número de nodos, en estas gráficas se puede observar que no existe ningún patrón obvio, por lo que se puede decir que los datos obtenidos mediante la función *DDC*(*G*) son independientes del tipo de red y del número de nodos existentes en la red.

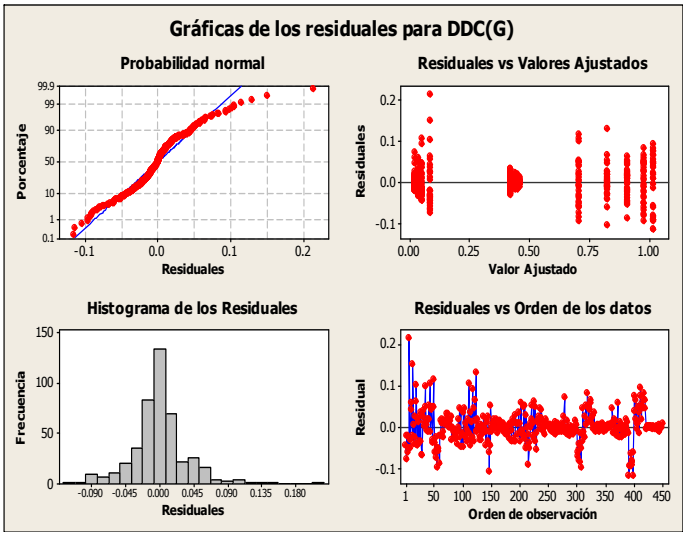
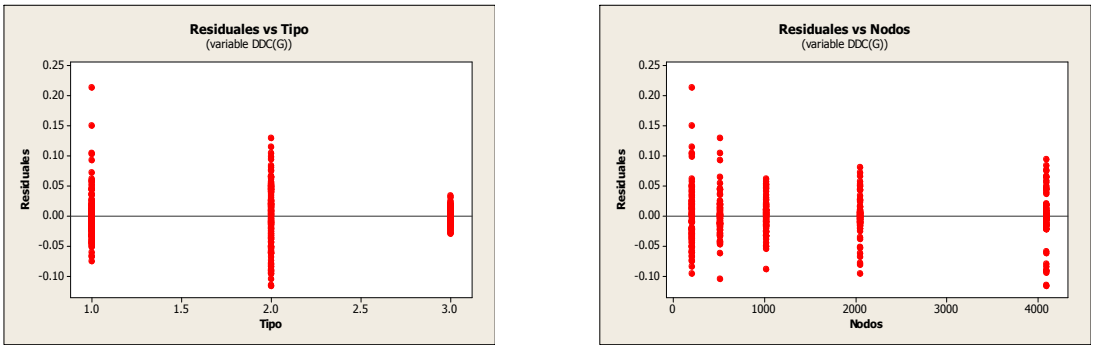


Figura 7.7 Gráfica de los residuales para la función *DDC*(*G*).



a) *b*)

Figura 7.8 Gráfica de los residuales para la función *DDC*(*G*) *a*) por tipo de red, *b*) por número de nodos

En el Anexo E, se pueden encontrar las gráficas de los residuales para las funciones de caracterización restantes y su discusión, es importante mencionar que las conclusiones para las funciones $\langle k \rangle$ y $CG(G)$ quedan descartadas por razones que se señalan en el Anexo E. En este punto se puede concluir que el factor tipo de red es el que tiene un mayor efecto sobre las funciones de caracterización, así también se detecto que las funciones de caracterización $\sigma_{\langle k \rangle}$, $DDC(G)$, $L(G)$, $D(G)$, E_{loc} y E_{glob} difieren significativamente de acuerdo al tipo de red y que el número de nodos tiene un efecto pequeño sobre estas funciones de caracterización a excepción de la función $\sigma_{\langle k \rangle}$.

De acuerdo a las conclusiones realizadas las funciones de caracterización relevantes son la desviación estándar del grado $\sigma_{\langle k \rangle}$, el coeficiente de dispersión del grado global $DDC(G)$, la longitud de ruta característica $L(G)$, el diámetro de la red $D(G)$, la eficiencia local E_{loc} y la eficiencia global E_{glob}.

Las funciones de caracterización irrelevantes son el grado promedio $\langle k \rangle$ y el coeficiente de agrupamiento $CG(G)$ ya que muestran anormalidades en las graficas de resiudales lo que hace descartar las conclusiones realizadas para estas funciones en el analisis de la varianza y el análisis de las graficas de los efectos.

7.2.2 Identificación de características muy relevantes y relevantes débiles

En este paso es de interés detectar diferencias específicas en los niveles del factor tipo de red que es el factor con mayor efecto y de esta manera identificar a qué casos (ver sección 6.3) pertenecen las funciones de caracterización y catalogarlas como características muy relevantes o relevantes débiles. Para lograr esto se realizaron comparaciones múltiples mediante la prueba de Tukey con un nivel de significación de $\alpha = 0.01$, resultados de estas pruebas se encuentran en el Cuadro f1 del Anexo F. Para este paso se utilizó el software estadístico SAS V8.

Observando los resultados, se tiene que las funciones de caracterización que difieren para cada tipo de red son $\sigma_{\langle k \rangle}$, *DDC*(*G*), *L*(*G*) y *D*(*G*), y representan al Caso 1, siendo estas características muy relevantes.

Por otro lado las funciones E_{loc} y E_{glob} son significativamente iguales para las redes Exponenciales (2) y las redes Power-Law (3), y representan al Caso 2, siendo estas características relevantes débiles.

7.2.3 Eliminación de características redundantes

Una vez que se han identificado las funciones de caracterización muy relevantes y relevantes débiles, es importante seleccionar funciones de caracterización que no estén correlacionadas, para de esta manera eliminar redundancia.

Para llevar a cabo esta selección se realizó un análisis de correlación de las funciones de caracterización muy relevantes y relevantes débiles, los resultados de este análisis se pueden observar en el Cuadro 7.1.

Cuadro 7.1 Resultados del análisis de correlación de las funciones de caracterización

Correlations: Std, DVC, L, D, EL, EG

	σ_<k>	DDC(G)	L(G)	D(G)	E_loc
DDC(G)	-0.240				
	0.000				

L(G)	-0.494	0.698			
	0.000	0.000			
D(G)	-0.552	0.696	**0.968**		
	0.000	0.000	0.000		
E_loc	0.108	-0.460	-0.559	-0.434	
	0.022	0.000	0.000	0.000	
E_glob	0.443	-0.759	**-0.944**	-0.897	0.760
	0.000	0.000	0.000	0.000	0.000

Cell Contents: Pearson correlation
P-Value

Se tomo como referencia para eliminar funciones de caracterización altamente correlacionadas, valores absolutos del coeficiente de Pearson mayores a 0.9. Primeramente se eliminaron aquellas funciones de caracterización relevantes débiles que estuvieran altamente correlacionadas con las funciones de caracterización muy relevantes, de esta manera la función E_{glob} se descartó por estar altamente correlacionada con la función $L(G)$ lo que representa redundancia. Posteriormente se pudo observar que la función $D(G)$ también esta altamente correlacionada con la función $L(G)$, tomándose la decisión de descartar a la función $D(G)$ ya que esta representa a la mayor de las distancias más cortas en la red y $L(G)$ representa el promedio de todas las distancias más cortas, por lo que esta es más representativa.

Fue así como las funciones de caracterización seleccionadas para posteriormente identificar el conjunto mínimo fueron $\sigma_{\langle k \rangle}$, $DDC(G)$, $L(G)$ y E_{loc}. De esta manera al menos una de las funciones de caracterización seleccionadas representa los casos 1 (características muy relevantes) y 2 (características relevantes débiles) mencionados en la sección 6.3.

7.2.4 Identificación del conjunto mínimo

En este paso el objetivo es obtener un subconjunto de funciones de caracterización mínimo, también llamado conjunto óptimo, que contenga la mayoría de las funciones de caracterización muy relevantes y algunas relevantes débiles, sin que haya redundancia entre ellas.

Para lograr este objetivo se realizó un análisis discriminante cuadrático con diferentes combinaciones de las funciones de caracterización identificadas como muy relevantes ($\sigma_{\langle k \rangle}$, $DDC(G)$, $L(G)$) y con las combinaciones de las funciones muy relevantes más la función de caracterización identificada como relevante débil (E_{loc}), de esta manera se puede observar que tan buenas son las funciones de caracterización para discriminar y si son necesarias todas las funciones de caracterización identificadas en la sección 7.2.3. El conjunto de entrenamiento y el conjunto de prueba fueron las redes generadas en la sección 7.1, se utilizó la técnica de validación cruzada [Witten 2005] para la fase de entrenamiento y la fase de prueba.

Para realizar el análisis discriminante cuadrático se empleo Minitab 14, en la Tabla 7.14 se muestran los resultados obtenidos con las diferentes combinaciones descritas en el párrafo anterior en la primera columna se muestra la combinación empleada en el análisis, de la segunda a la cuarta columna se muestra el porcentaje y número de instancias de red clasificadas correctamente en cada tipo de red, y la última columna muestra el porcentaje total de las redes clasificadas correctamente.

Tabla 7.14 Resultados del Análisis Discriminante Cuadrático.

Análisis Discriminante Cuadrático				
Combinación de funciones de caracterización	**Redes Aleatorias**	**Redes Power-Law**	**Redes Exponenciales**	**% Total**
$\sigma_{\langle k \rangle}$	44% (66/150)	68% (102/150)	87.3% (131/150)	66.7%
***DDC*(*G*)**	**99.3% (149/150)**	**100% (150/150)**	**100% (150/150)**	**99.8%**
$L(G)$	98.7% (148/150)	67.3% (101/150)	48% (72/150)	71.3%
$\sigma_{\langle k \rangle}$, $DDC(G)$	99.3% (149/150)	100% (150/150)	100% (150/150)	99.8%
$\sigma_{\langle k \rangle}$, $L(G)$	97.3% (146/150)	82% (123/150)	86.7% (130/150)	88.7%
$DDC(G)$, $L(G)$	99.3% (149/150)	100% (150/150)	100% (150/150)	99.8%
$\sigma_{\langle k \rangle}$, *DDC*(*G*), *L*(*G*)	**100% (150/150)**	**100% (150/150)**	**100% (150/150)**	**100%**

$\sigma_{\langle k \rangle}$, E_{loc}	91.3% (137/150)	100% (150/150)	100% (150/150)	97.1%
$DDC(G)$, E_{loc}	99.3% (149/150)	100% (150/150)	100% (150/150)	99.8%
$L(G)$, E_{loc}	98.7% (148/150)	60% (90/150)	73.3% (110/150)	77.8%
$\sigma_{\langle k \rangle}$, $DDC(G)$, E_{loc}	99.3% (149/150)	100% (150/150)	100% (150/150)	99.8%
$\sigma_{\langle k \rangle}$, $L(G)$, E_{loc}	99.3% (149/150)	100% (150/150)	100% (150/150)	99.8%
$DDC(G)$, $L(G)$, E_{loc}	99.3% (149/150)	100% (150/150)	100% (150/150)	99.8%
$\sigma_{\langle k \rangle}$, $DDC(G)$, $L(G)$, E_{loc}	**99.3% (149/150)**	**100% (150/150)**	**100% (150/150)**	**99.8%**

Se puede observar que con las funciones de caracterización seleccionadas en la sección 7.2.3 ($\sigma_{\langle k \rangle}$, $DDC(G)$, $L(G)$, E_{loc}) se obtiene un porcentaje total del 99.8% de acierto, con estas funciones sólo una instancia de tipo aleatoria no se clasifica como tal.

De las características identificadas como muy relevantes es la función $DDC(G)$ que por si sola tiene mejor capacidad para distinguir entre los diferentes tipos de red al obtener un porcentaje total del 99.8% con esta función solo una instancia de tipo aleatoria no se clasifica como tal.

Las funciones de caracterización con las cuales se obtiene un porcentaje total del 100% de acierto son $\sigma_{\langle k \rangle}$, $DDC(G)$, $L(G)$, por lo que se puede concluir que la función E_{loc} identificada como relevante débil no contribuye a mejorar los resultados del desempeño del análisis discriminante.

Se puede concluir que el conjunto mínimo F' contendrá la función $DDC(G)$, una razón para descartar la función $\sigma_{\langle k \rangle}$ es que en el análisis estadístico se observa que esta función se ve afectada considerablemente por el número de nodos existentes en la red, y la función $L(G)$ requiere mucho tiempo de computo para calcularse, por lo que $F' = DDC(G)$ será el conjunto óptimo que satisface $P(C \mid F' = \vec{f}') @ P(C \mid F = \vec{f})$.

Utilizando un algoritmo de clasificación basado en un árbol de decisión como el C4.5 (implementación en Weka J48 [Servente 2002]) se puede observar en la Figura 7.9 a la

función *DDC*(*G*) como única característica para clasificar instancias de red, con esto podemos concluir una vez más que la función de caracterización con mejores características para discriminar entre las redes Aleatorias, redes Power-Law y redes Exponenciales es la función de caracterización *DDC*(*G*).

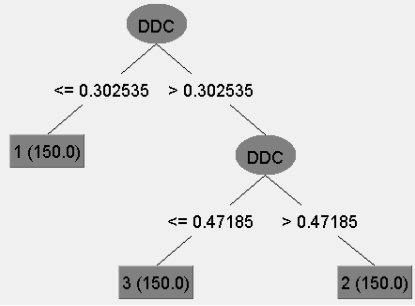


Figura 7.9 Árbol de decisión obtenido con el algoritmo C4.5

En esta etapa de selección de características se hizo uso de diferentes pruebas estadísticas que permitieron identificar las funciones de caracterización relevantes y no redundantes que forman el conjunto mínimo, en la Tabla 7.15 se muestra las herramientas estadísticas que se utilizaron en cada actividad de la etapa de selección de características y el resultado obtenido.

Tabla 7.15 Herramientas utilizadas en la selección de características.

Selección de características		
Actividad	**Herramienta Estadística**	**Resultado**
Identificación de características relevantes e irrelevantes	Diseño Experimental: Factorial. Pruebas Estadísticas: Análisis Multivariado de la Varianza, Análisis de Residuales, Gráficas de interacción.	✓ Características relevantes: $\sigma_{\langle k \rangle}$, $DDC(G)$, $L(G)$, $D(G)$, E_{loc}, E_{glob} × Características irrelevantes: $\langle k \rangle$, $CG(G)$
Identificación de características muy relevantes y relevantes débiles	Comparaciones múltiples usando la Prueba de Tukey	Características muy relevantes: $\sigma_{\langle k \rangle}$, $DDC(G)$, $L(G)$, $D(G)$ Características relevantes débiles: E_{loc}, E_{glob}

Eliminación de redundancia	Análisis de Correlación	✓ Características no redundantes: $\sigma_{\langle k \rangle}$, $DDC(G)$, $L(G)$, E_{loc} × Características redundantes: $D(G)$, E_{glob}
Identificación del conjunto mínimo	Análisis Discriminante	Conjunto mínimo: $DDC(G)$.

7.3 CLASIFICACIÓN DE REDES COMPLEJAS

Una vez que se identificó el subconjunto mínimo $F' = \{DDC(G)\}$, esta función fue utilizada para determinar el clasificador con mejor desempeño y clasificar instancias de red de naturaleza desconocida, para realizar esta actividad se utilizo Weka una colección de algoritmos de aprendizaje automático y herramientas de preprocesamiento para minería de datos [Witten 2005].

7.3.1 Identificación del clasificador con el mejor desempeño

En este paso se tomaron como base los algoritmos de clasificación utilizados en el estado del arte y mediante la interfaz de Weka llamada *Experimenter* se compararon los algoritmos de clasificación Naive Bayes, J48, SimpleLogistic, Multilayer Perceptron y BayesNet. El objetivo en este paso es determinar el algoritmo de clasificación que tiene mejor desempeño utilizando como entrada la función de caracterización $DDC(G)$.

Cuadro 7.2 Comparación de algoritmos de clasificación que usan como entrada la función de caracterización *DDC*(*G*)

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Percent_correct
Datasets:   1
Resultsets: 5
Confidence: 0.01 (two tailed)
Sorted by:  -
Date:       1/07/07 09:52 PM


Dataset                   (1) bayes.Na | (2) trees (3) funct (4) funct (5) bayes
--------------------------------------------------------------------------------
funcion_DDC(G)           (100)   99.78 |   99.49     99.78     99.78     99.93
--------------------------------------------------------------------------------
```

```
                                (v/ /*) |   (0/1/0)    (0/1/0)    (0/1/0)    (0/1/0)


Key:
(1) bayes.NaiveBayes '' 5995231201785697655
(2) trees.J48 '-C 0.25 -M 2' -217733168393644444
(3) functions.SimpleLogistic '-I 0 -M 500 -H 50 -W 0.0' 7397710626304705059
(4) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a'
572250905027665169
(5) bayes.BayesNet '-D -Q bayes.net.search.local.K2 -- -P 1 -S BAYES -E
bayes.net.estimate.SimpleEstimator -- -A 0.5' 746037443258775954
```

En el Cuadro 7.2 se muestran los resultados de la comparación de los algoritmos de clasificación mencionados en el párrafo anterior, el parámetro analizado fue el porcentaje de instancias de red clasificadas correctamente, el algoritmo base con el cual se van a comparar los demás algoritmos es el Naive Bayes, se puede observar que los otros cuatro algoritmos tienen debajo (0/1/0) esto indica que el desempeño de los algoritmos J48, SimpleLogistic, Multilayer Perceptron y BayesNet es significativamente igual al desempeño del algoritmo Naive Bayes considerando un nivel de confianza de $\alpha = 0.01$. [Witten 2005].

De acuerdo con los resultados obtenidos mediante la comparación de algoritmos se tomó la decisión de utilizar el algoritmo Naive Bayes, para clasificar instancias de naturaleza desconocida e identificar el tipo de red en cada punto de instancias de Internet. Para ello se entreno al clasificador Naive Bayes con las instancias de red generadas, de esta manera se creó un modelo (véase Figura 6.1) con el cual se realizarán las clasificaciones.

7.3.2 Clasificación de instancias de naturaleza desconocida

Para llevar a cabo esta actividad se estableció contacto con otros investigadores del área de redes complejas para solicitar instancias de red de las cuales no se conoce su tipo. Se solicitaron instancias de red con una cantidad de nodos de entre 200 y 4096 nodos.

A estas instancias de red se les aplicó la función de caracterización que pertenece al conjunto óptimo $F' = \{DDC(G)\}$ y usando el modelo que se obtuvo en la sección 7.3.1, se clasificaron 160 instancias de red de naturaleza desconocida, los resultados de la clasificación se muestran en el Anexo G.

Posteriormente se hizo de nuestro conocimiento los modelos empleados en la generación de instancias de naturaleza desconocida y una relación de qué instancia fue generada con qué modelo, los modelos utilizados fueron Caveman, Watts - Strogratz, Kleinberg una descripción de estos modelos puede encontrarse en [Virtanen 2003]; también se utilizaron los modelos Erdos y Renyi, Barabási-Albert y el modelo Geographical Scale Free Graphs [Fabrikant 2002]. Los modelos Caveman, Watts - Strogratz, Kleinberg y Erdös y Rényi reproducen redes Aleatorias, los modelos Geographical Scale Free Graphs y Barabási-Albert reproducen redes Power-Law.

Con esta información se compararon los resultados arrojados por el clasificador, se tiene que, de 160 instancias de red, se clasificaron correctamente 146 instancias, es decir, un 91.25% de las instancias. Se puede concluir que la función coeficiente de dispersión del grado *DDC*(*G*) puede identificar cuantitativamente tipo de red con una eficacia aceptable.

7.3.3 Determinación del tipo de red en cada nodo de las instancias de Internet

En este paso se tomaron instancias de Internet disponibles en [CAIDA 2005], estas instancias reflejan la topología de Internet a nivel de sistemas autónomos (SA) para los años de 1997, 2000 y 2003.

En la Tabla 7.16 se muestra información acerca de las instancias de Internet, en la primera columna se muestra el nombre de la instancia, la segunda y tercera columna muestran el números de nodos y el número de aristas respectivamente, cuarta columna el grado promedio de la red, en la quinta y sexta columna se muestra el grado mínimo y el grado máximo respectivamente y en la última columna se muestra el coeficiente de dispersión del grado global.

De acuerdo con diversas investigaciones [Albert 2000, Faloutsos 1999] se puede afirmar que Internet tiene en promedio distribución del grado Power-Law, eso se reafirma con los valores calculados para las instancias de Internet descritas anteriormente y el árbol de decisión mostrado en la Figura 7.9.

Tabla 7.16 Información de las instancias de Interner para los años 1997, 2000 y 2003

Instancia	*N*	*e*	$\langle k \rangle$	$\delta(G)$	$\Delta(G)$	*DDC(G)*
INT – 1997	3015	5,156	3.42	1	590	1.02
INT – 2000	6474	12572	3.88	1	1458	1.06
INT – 2003	192244	609066	6.33	1	1071	0.75

Hasta este punto la función de caracterización usada es el $DDC(G)$ el coeficiente de dispersión del grado global, de manera que el modelo construido por el algoritmo de clasificación ha capturado a través de esta función el comportamiento global de las redes estudiadas.

Como se muestra en la Figura 7.10, ahora aplicaremos la función coeficiente de dispersión del grado a cada subgrafo de la red $DDC(G_i)$, el cual se tomará como un todo, de esta manera se podrá distinguir el tipo de red localmente. Cabe recordar que cada nodo tiene asociado un subgrafo en la red.

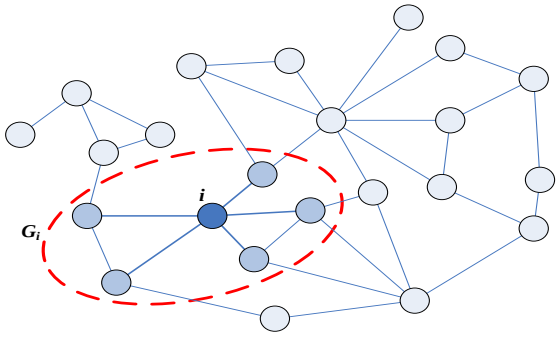


Figura 7.10 Identificación del tipo de red localmente

En la instancia de Internet que corresponde al año 1997 de los 3015 subgrafos, 2164 que corresponde al 72% de los subgrafos son de tipo Power-Law, 706 subgrafos que representa el 23% de los existentes son de tipo aleatorios y 145 subgrafos que representan el 5 % del total son de tipo Exponenciales, estos resultados se muestran en la Figura 7.11.

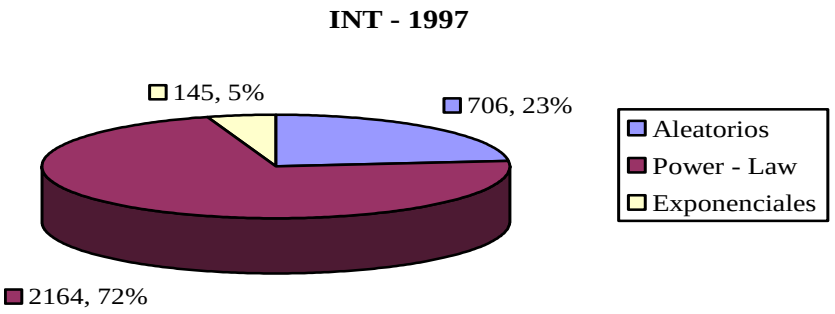


Figura 7.11 Tipo de red de los subgrafos de la instancia INT – 1997

En la instancia de Internet que corresponde al año 2000 de los 6474 subgrafos 4426 que corresponde al 68% de los subgrafos son de tipo Power-Law, 1134 subgrafos que representa el 18% de los existentes son de tipo aleatorios y 914 subgrafos que representan el 14 % del total son de tipo Exponenciales, estos resultados se muestran en la Figura 7.12.

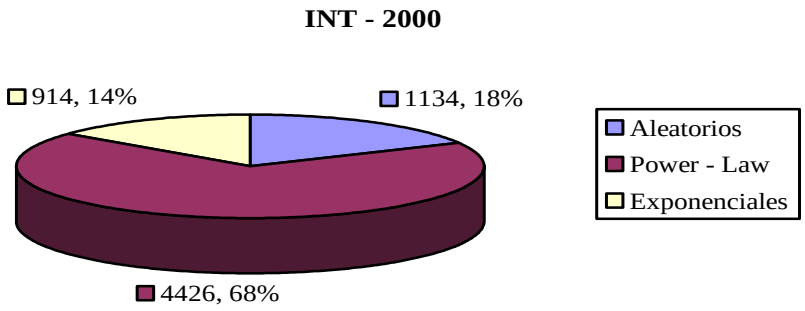


Figura 7.12 Tipo de red de los subgrafos de la instancia INT – 2000

En la instancias de Internet que corresponde al año 2003 de los 192244 subgrafos 139150 que corresponde al 73% de los subgrafos son de tipo Power-Law, 29717 subgrafos que representa el 15% de los existentes son de tipo aleatorios y 23377 subgrafos que representan el 12 % del total son de tipo Exponenciales, estos resultados se muestran en la Figura 7.13.

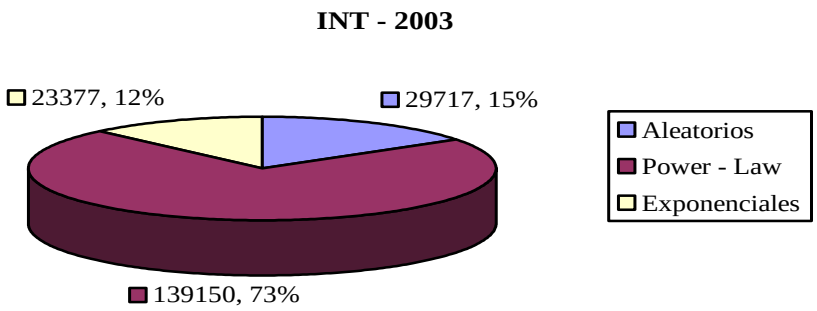


Figura 7.13 Tipo de red de los subgrafos de la instancia INT – 2003

Capítulo 8

CONCLUSIONES Y TRABAJOS FUTUROS

En este capitulo se presentan las aportaciones de esta investigación y se sugieren algunos trabajos futuros.

8.1 CONCLUSIONES

En el presente trabajo de investigación se desarrolló una metodología que toma como base la arquitectura de un agente de aprendizaje, para identificar el conjunto mínimo de funciones de caracterización que permitan discriminar cuantitativamente entre diferentes tipos de redes complejas como las redes Aleatorias, Power-Law y Exponenciales.

Para cumplir con el objetivo planteado se realizó un diseño experimental en el cual se utilizaron un conjunto de herramientas estadísticas que permitieron seleccionar funciones de caracterización cuantitativas relevantes y no redundantes e identificar aquellas funciones que forman el conjunto mínimo F'.

Los resultados obtenidos mediante el enfoque estadístico empleado para la selección de características relevantes y no redundantes se compararon con resultados de diversos métodos y algoritmos usados ampliamente dentro del área de Análisis Multivariado y Aprendizaje Automático para la selección de características (ver Anexo H) tales como: selección hacia delante, eliminación hacia atrás, algoritmo RELIEF, entre otros.

Si bien los resultados fueron los mismos, el enfoque estadístico empleado permite explicar el por qué ciertas funciones del caracterización fueron seleccionas y otras fueron rechazadas, y cómo es que los factores considerados en el diseño experimental influyen en las funciones de caracterización. Así también ayuda a explicar de cierta manera porqué los algoritmos de clasificación empleados tienen un buen desempeño usando el conjunto mínimo como entrada.

Los resultados de esta investigación muestran que la función de caracterización coeficiente de dispersión del grado global *DDC(G*) permite de manera cuantitativa discriminar entre redes Aleatorias, redes Power-Law y redes Exponenciales.

El Análisis Discriminante Cuadrático efectuado sobre las instancias generadas usando como variable de entrada el *DDC*(*G*) obtuvo una exactitud de clasificación del 99.8% con las instancias generadas para el diseño experimental, el algoritmo Naive Bayes pudo identificar correctamente el 99.78% de las instancias del diseño experimental y un 91.25% de instancias de naturaleza desconocida generadas por modelos diferentes a los considerados en esta investigación.

Se pudo observar que el tipo de red es el factor que influye de manera importante en los valores obtenidos por la función de caracterización *DDC(G)* y no la cantidad de nodos presentes en la red, se concluye que la función de caracterización *DDC(G)* tiende a ciertos valores dependiendo del tipo de red:

- ✓ En la redes Aleatorias $DDC(G) \to 0$, esto debido a que la mayoría de los nodos tienen aproximadamente el mismo grado, cercano al grado promedio de la red $\langle k \rangle$, se puede decir que la dispersión del grado en promedio es nula o muy pequeña.
- ✓ En la redes Exponenciales $DDC(G) \to 0.4$, esto debido a que la mayoría de los nodos tienen un grado cercano al grado promedio de la red $\langle k \rangle$ y se pueden observar un número muy reducido de nodos con un alto grado por arriba de $\langle k \rangle$.

- ✓ En la redes Power-Law $DDC(G) \rightarrow 1$, esto debido a que solo unos cuantos nodos poseen un grado muy alto y la mayoría de los nodos un grado muy pequeño, por lo que la dispersión del grado en promedio tiende a ser alta.

Estas observaciones permiten concluir que el modelo de aprendizaje construido por el algoritmo de clasificación basado en la función de caracterización *DDC*(*G*) captura el comportamiento global de estos tipos de redes, por lo que si se calcula la función de caracterización *DDC* sobre un subgrafo de una red y se aplica el modelo construido se puede distinguir el tipo de red localmente.

Otro punto importante de mencionar es que en esta investigación el número de funciones de caracterización utilizadas para la discriminación de los tipos de redes, es muy pequeño en relación con el número de funciones utilizadas en los trabajos relacionados descritos en el estado del arte, esto reduce dramáticamente el tiempo de computo requerido para la clasificación de redes complejas.

Así mismo se reafirma que fijar el número de aristas de los grafos, es decir, que los grafos tengan el mismo grado promedio, ayuda a comprender fenómenos de interés pero introduce peligros al momento de realizar pruebas estadísticas, ya que no hay variabilidad en los datos.

Debido a que el tiempo de cómputo requerido para generar y caracterizar las instancias de red se incrementa significativamente de acuerdo con el tamaño de la red, un aspecto importante en este trabajo, fue determinar estadísticamente el número mínimo de instancias a generar de manera que se pudieran detectar diferencias significativas en los experimentos llevados a cabo en esta investigación.

8.2 TRABAJOS FUTUROS

Los resultados de esta investigación muestran que la función de caracterización coeficiente de dispersión de grado (*DDC*) permite distinguir cuantitativamente entre diferentes redes complejas como las Aleatorias, Power-Law y Exponenciales. Sin embargo del análisis de

los resultados obtenidos en el proceso de clasificación se derivan cuestiones importantes que se sugieren abordar en trabajos futuros:

a) Identificar las estructuras de red que son mal clasificadas por el algoritmo de clasificación basado en la función de caracterización *DDC*, explicar el por qué esas estructuras son mal clasificadas y analizar que es lo que tienen en común. Este punto es primordial ya que no permitirá identificar los rangos en el que el *DDC* no pueda decir con exactitud que tipo de red es, o no permitan identificar si una estructura.

b) Realizar experimentaciones con instancias de sistemas naturales modelados como redes complejas.

ANEXO A

A1. INSTANCIAS DE REDES REALES

Con el objetivo de promover el estudio de las redes complejas, diversas organizaciones ponen a disposición herramientas e instancias de redes, a continuación mencionaremos algunas organizaciones que se dan a esta tarea:

CAIDA (Cooperative Association for Internet Data Análisis): Es una iniciativa de organizaciones comerciales, gubernamentales y de investigación, que tiene como objetivo promover la cooperación en la ingeniería y mantenimiento robusto, escalable y global de la infraestructura de Internet, para ello provee herramientas, datos y diversos análisis realizados a Internet. CAIDA promueve la medición macroscópica y análisis de Internet y su desempeño, colecciona datos para análisis del funcionamiento y pone a disposición de la comunidad científica diferentes tipos de datos referentes a Internet. Los datos están agrupados en diferentes categorías como la topología, seguridad, enrutamiento, entre otras [CAIDA 2005].

En nuestro caso el tópico de interés es la topología, los datos que se encuentran en esta categoría describen la infraestructura de enlaces de la red a varios niveles y los cuales tienen como objetivo el estudio de las interconexiones entre dispositivos que componen Internet, en nuestro caso nos interesa los siguientes datos:

- CAIDA's Macroscopic Topology AS Adjacencies: contiene la matriz de adyacencia del grafo de Internet (nivel de Sistemas Autónomos), esta información se obtiene de las tablas BGP.
- CAIDA's Router-Level Topology Measurements: contiene la matriz de adyacencia del grafo asociado a Internet (nivel de ruteadores), aquí se tienen dos versiones de esta matriz una de ellas es la matriz del grafo dirigido y la otra es la del grafo no dirigido, siendo de interés para este trabajo el grafo no dirigido.

CASOS (Center of Computational Analysis of Social and Organizational Systems): Tiene como objetivo el estudio de redes complejas socio-tecnológicas con el fin de comprender los principios organizacionales, de coordinación, administración y desestabilización de sistemas de agentes adaptables inteligentes comprometidos en tareas de equipo reales a nivel organizacional o social.

CASOS pone a disposición información de diferentes tipos de redes sociales con su correspondiente análisis topológico [CASOS 2005].

Network Workbench: Esta comunidad tiene como objetivo dar soporte a las investigaciones de la ciencia de las redes en los campos de la biomedicina, las Ciencias Sociales y la Física. Aunque en fase de construcción esta comunidad comparte bases de datos acerca de redes, sobre las cuales se pueden realizar procesos de análisis y visualizaciones, en un futuro no solo compartirá bases de datos de redes sino también herramientas para la visualización de redes y la interacción con estas así como herramientas para la generación, ejecución y validación de modelos de redes para avanzar en la comprensión de la estructura y dinámica de las redes complejas [NWB 2005].

En este sitio se podrá encontrar instancias de redes del biológicas, económicas, sociales, científicas y de Internet; algunas instancias ya están disponibles.

ANEXO B

En este anexo se muestran los estadísticos descriptivos e histogramas de las funciones de caracterización por tipo de red y por número de nodos ajustados a la distribución normal, obtenidos a través del software estadístico Minitab 14. Los histogramas son usados para examinar la forma y la dispersión de los datos, también pueden mostrar cierta evidencia acerca de la semejanza en la distribución de dos o más poblaciones, en este trabajo los datos pueden representarse en poblaciones por tipo de red no importando en número de nodos existentes en la red y poblaciones por número de nodos no importando el tipo de red. El histograma puede representarse como el ajuste de los datos a la distribución normal representada por una curva.

B1. Histogramas de las funciones de caracterización según el tipo de red

En el Cuadro b1 se muestran algunos estadísticos descriptivos, como la media, la desviación estándar, el valor máximo y mínimo de las funciones de caracterización $\langle k \rangle, \sigma_{\langle k \rangle}, L(G), D(G), E_{loc}, E_{glob}, C(G)$ según el tipo de red.

En la función $\langle k \rangle$ se puede observar que las medias de las redes Power-Law (2) y Exponenciales (3) son similares y la media de las redes Aleatorias (1) es diferente a la de las Power-Law (2) y Exponenciales (1), pero los valores de la desviación estándar y los valores máximos y mínimos dan evidencias de un traslape entre las redes, lo que se puede observar en la Figura b1.

En la función $\sigma_{\langle k \rangle}$ se puede observar que las medias de los tres tipos de redes son diferentes, pero la desviación estándar, los valores máximos y mínimos muestran indicios de traslape entre las redes, lo que puede observarse en la Figura b2.

Las funciones $L(G)$ y $D(G)$ tienen un comportamiento similar, las medias de las redes Power-Law (2) y Exponenciales (3) son similares y la media de las redes Aleatorias (1) es diferente a la de las redes Power-Law (2) y Exponenciales (3), pero los valores de la desviación estándar y los valores máximos y mínimos dan evidencias de un traslape entre las redes, lo que se puede observar en la Figura b3 y la Figura b4.

En la función E_{loc} se puede observar que las medias de las redes Power-Law (2) y Exponenciales (3) son similares y la media de las redes Aleatorias (1) es diferente a la de las Power-Law y Exponenciales, pero los valores de la desviación estándar y los valores máximos y mínimos dan evidencias de un traslape entre las redes, lo que se puede observar en la Figura b5. La función E_{glob} tiene un comportamiento similar a la función E_{loc} solo que la media de las redes Aleatorias (1) tiende a alejarse de las redes Power-Law (2) y Exponenciales (3), esto se puede observar en la Figura b6.

En la función $CG(G)$ se puede observar que las medias de de las redes Power-Law (2) y Exponenciales (3) tienden a ser similares y la media de las redes Aleatorias (1) es diferente a la de las redes Power-Law (2) y Exponenciales (3), pero los valores de la desviación estándar y los valores máximos y mínimos dan evidencias de un traslape entre las redes, lo que se puede observar en la Figura b7.

Cuadro b1. Estadísticos descriptivos de las funciones de caracterización de acuerdo al tipo de red no importando el número de nodos.

Descriptive Statistics: <k>, σ_<k>, DDC(G), L(G), D(G), E_loc, E_glob, CG(G)

Variable	Tipo	Total Count	Mean	StDev	Minimum	Maximum
<k>	1	150	764.6	888.1	7.00	4058.0
	2	150	9.800	3.301	5.000	14.000
	3	150	9.560	3.195	5.000	14.000
σ_<k>	1	150	14.171	8.003	2.543	32.073
	2	150	10.033	2.964	5.108	16.076
	3	150	5.092	1.471	3.098	7.566
L(G)	1	150	1.5044	0.3123	1.0005	2.8585
	2	150	3.0895	0.4801	2.2529	3.9201
	3	150	3.3987	0.6197	2.3185	4.5836
D(G)	1	150	2.1067	0.3868	2.0000	5.0000
	2	150	4.9333	0.8798	3.0000	7.0000
	3	150	5.3800	1.0661	4.0000	8.0000

Continuación Cuadro b1. Estadísticos descriptivos de las funciones de caracterización de acuerdo al tipo de red no importando el número de nodos.

E_loc	1	150	0.7549	0.1349	0.5238	0.9955
	2	150	0.65112	0.03900	0.59474	0.72060
	3	150	0.63034	0.03455	0.58175	0.68588
E_glob	1	150	0.7483	0.1432	0.3829	0.9955
	2	150	0.35495	0.05674	0.26857	0.48056
	3	150	0.32627	0.06496	0.22848	0.46970
CG(G)	1	150	0.5003	0.2793	0.0310	0.9910
	2	150	0.05300	0.04083	0.00934	0.15828
	3	150	0.02248	0.02406	0.00164	0.09566

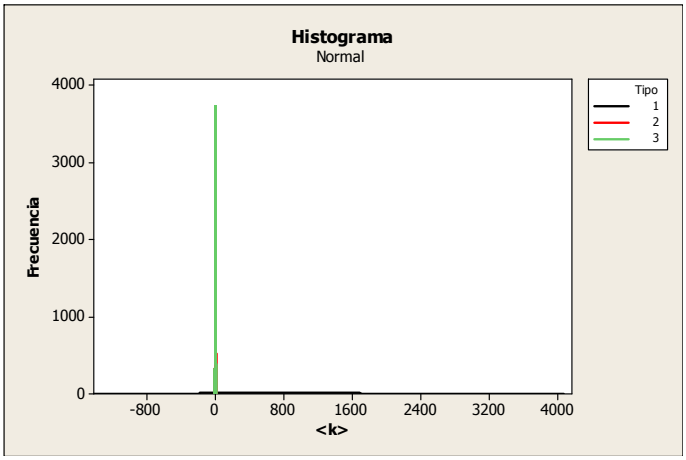


Figura b1. Histograma de la función $\langle k \rangle$ por tipo de red.

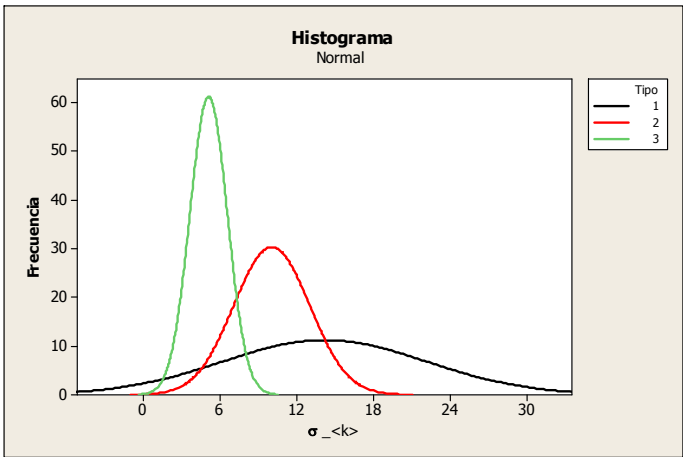


Figura b2. Histograma de la función $\sigma_{\langle k \rangle}$ por tipo de red.

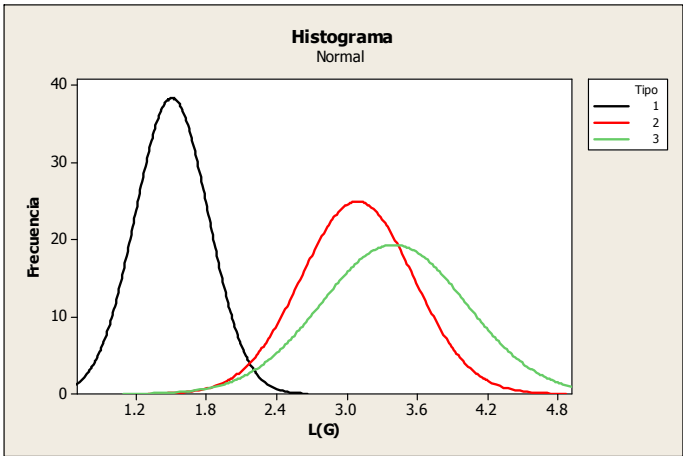


Figura b3. Histograma de la función *L*(*G*) por tipo de red.

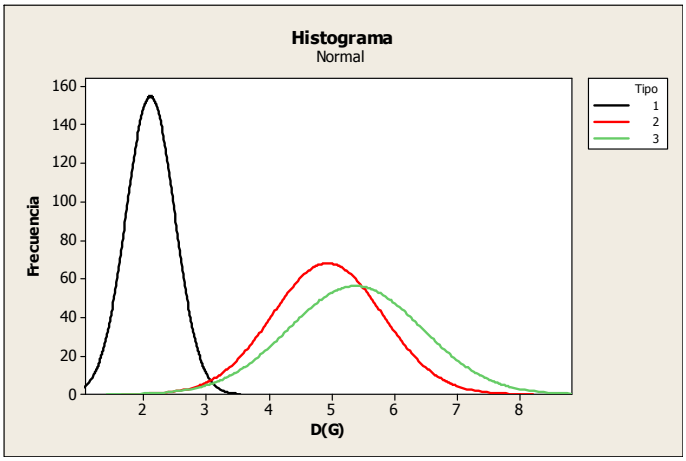


Figura b4. Histograma de la función *D*(*G*) por tipo de red.

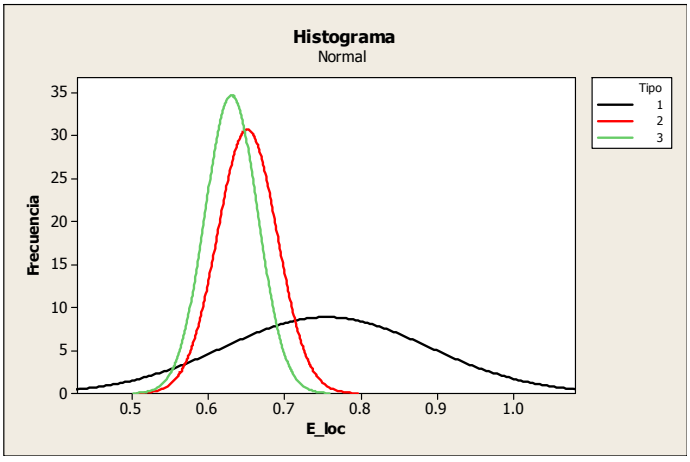


Figura b5. Histograma de la función E_{loc} por tipo de red.

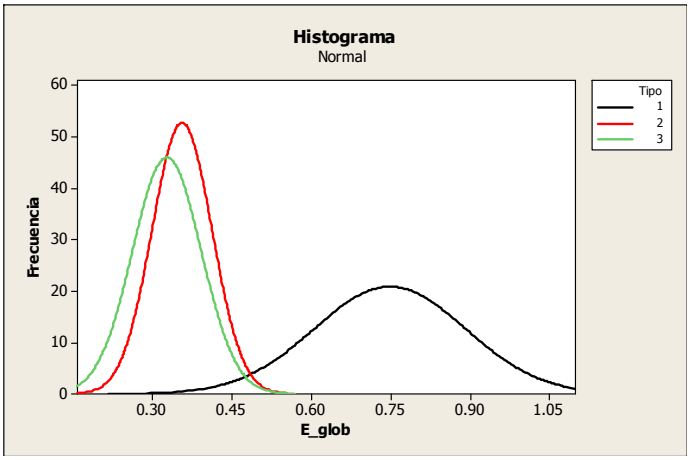


Figura b6. Histograma de la función E_{glob} por tipo de red.

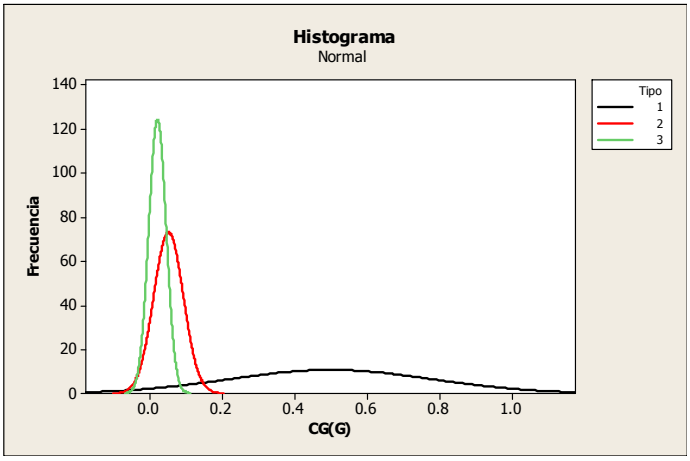


Figura b7. Histograma de la función *CG*(*G*) por tipo de red.

B2. Histogramas de las funciones de caracterización según el número de nodos

En el Cuadro b2 se muestran algunos estadísticos descriptivos, como la media, la desviación estándar, el valor máximo y mínimo de las funciones de caracterización $\langle k \rangle, \sigma_{\langle k \rangle}, L(G), D(G), E_{loc}, E_{glob}, C(G)$ según el número de nodos existentes en la red.

En las funciones $\langle k \rangle$ y $\sigma_{\langle k \rangle}$ se puede observar que las medias de las redes según el número de nodos son diferentes, los valores de la desviación estándar y los valores máximos y mínimos dan evidencias de un traslape entre las redes con diferente número de nodos, lo que se puede observar en la Figura b8 y en la Figura b9 respectivamente.

Cuadro b2. Estadísticos descriptivos de las funciones de caracterización de acuerdo al número de nodos existentes en la red no importando el tipo de red.

Descriptive Statistics: <k>, σ_<k>, DDC(G), L(G), D(G), E_loc, E_glob, CG(G)

Variable	Nodos	Total Count	Mean	StDev	Minimum	Maximum
<k>	200	90	38.91	52.63	5.00	191.00
	512	90	93.1	144.6	6.00	497.0
	1024	90	176.6	295.5	6.00	989.0
	2048	90	380.9	592.9	6.00	2022.0
	4096	90	617	1127	6.00	4058
σ_ <k>	200	90	6.067	1.893	2.543	9.922
	512	90	7.789	2.740	3.210	12.556
	1024	90	9.155	4.170	3.311	16.044
	2048	90	12.040	6.588	3.361	22.874
	4096	90	13.776	9.147	3.463	32.073
DDC(G)	200	90	0.4057	0.2595	0.0133	0.8235
	512	90	0.4413	0.3194	0.00766	0.9584
	1024	90	0.4670	0.3588	0.00561	0.9718
	2048	90	0.4858	0.3945	0.00247	1.0616
	4096	90	0.5011	0.4111	0.00151	1.1118
L(G)	200	90	2.2570	0.6056	1.0393	3.1583
	512	90	2.4788	0.7771	1.0271	3.6523
	1024	90	2.7007	0.9351	1.0331	4.0153
	2048	90	2.804	1.041	1.012	4.331
	4096	90	3.080	1.160	1.001	4.584
D(G)	200	90	3.633	1.126	2.000	5.000
	512	90	3.822	1.411	2.000	6.000
	1024	90	4.233	1.683	2.000	7.000
	2048	90	4.378	1.796	2.000	7.000
	4096	90	4.633	2.019	2.000	8.000
E_loc	200	90	0.69480	0.08819	0.59574	0.98053
	512	90	0.6807	0.1008	0.5606	0.9865
	1024	90	0.6780	0.1036	0.5563	0.9835
	2048	90	0.6794	0.0997	0.5344	0.9938
	4096	90	0.6609	0.1040	0.5238	0.9955
E_glob	200	90	0.5292	0.1781	0.3493	0.9804
	512	90	0.4940	0.2043	0.2952	0.9865
	1024	90	0.4666	0.2211	0.2659	0.9835
	2048	90	0.4609	0.2344	0.2384	0.9938
	4096	90	0.4318	0.2270	0.2285	0.9955
CG(G)	200	90	0.2285	0.2490	0.0273	0.9606
	512	90	0.1994	0.2736	0.0110	0.9729
	1024	90	0.1836	0.2829	0.00688	0.9669
	2048	90	0.1933	0.2852	0.00307	0.9876
	4096	90	0.1549	0.2729	0.00164	0.9910

Las funciones $L(G)$ y $D(G)$ tienen un comportamiento similar, en estas funciones se puede observar que las medias difieren según el número de nodos, aunque no tan drásticamente como las funciones $\langle k \rangle$ y $\sigma_{\langle k \rangle}$, los valores de la desviación estándar y los

valores máximos y mínimos dan evidencias de un traslape entre las redes con diferente número de nodos, lo que se puede observar en la Figura b10 y en la Figura b11 respectivamente.

En la función E_{loc} se puede observar que las medias de las redes según el número de nodos son similares, los valores de la desviación estándar y los valores máximos y mínimos dan evidencias de un traslape entre las redes con diferente número de nodos, lo que se puede observar en la Figura b12.

En la función E_{glob} se puede observar que las medias de las redes según el número de nodos difieren un poco, los valores de la desviación estándar y los valores máximos y mínimos dan evidencias de un traslape entre las redes con diferente número de nodos, lo que se puede observar en la Figura b13.

La función $CG(G)$ tiene un comportamiento similar a la función E_{loc}, donde se puede observar que las medias de las redes según el número de nodos son similares, los valores de la desviación estándar y los valores máximos y mínimos dan evidencias de un traslape entre las redes con diferente número de nodos, lo que se puede observar en la Figura b14.

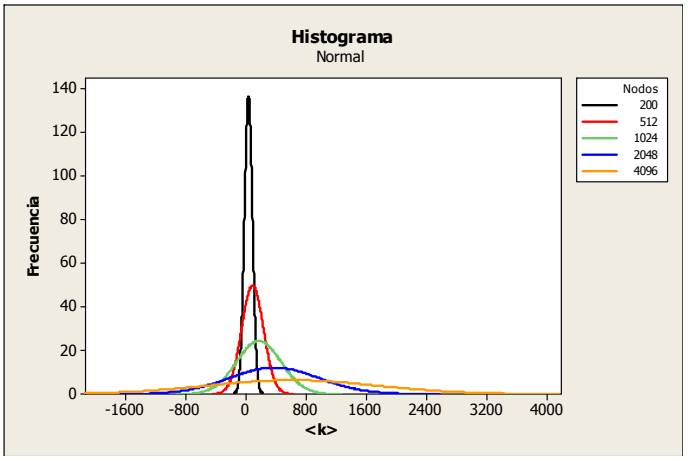


Figura b8. Histograma de la función $\langle k \rangle$ por número de nodos.

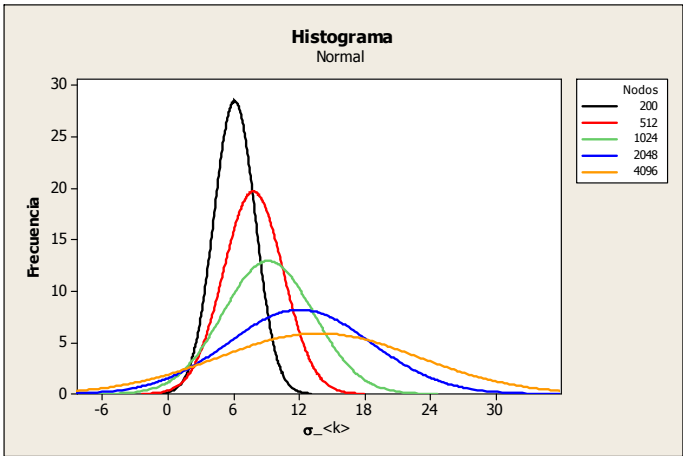


Figura b9. Histograma de la función $\sigma_{\langle k \rangle}$ por cantidad de nodos.

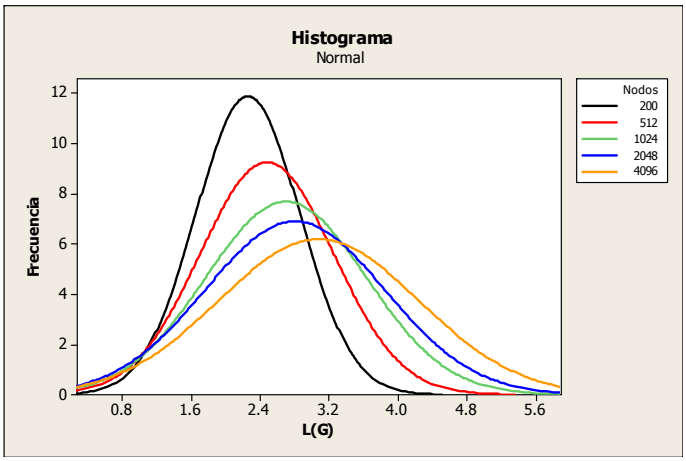


Figura b10. Histograma de la longitud de la función *L*(*G*) por cantidad de nodos.

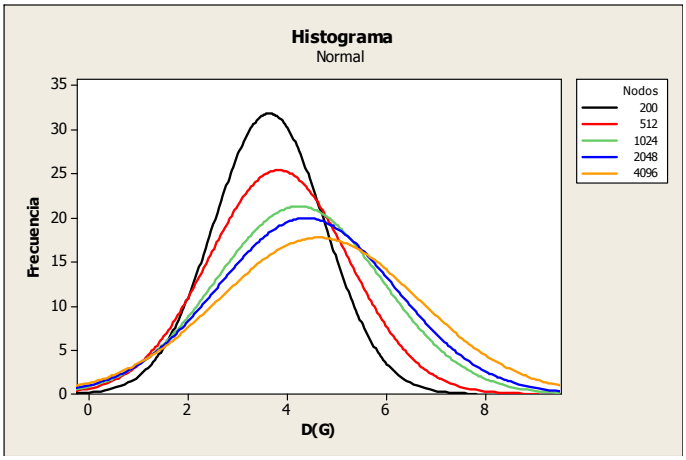


Figura b11. Histograma de la función *D*(*G*) por cantidad de nodos.

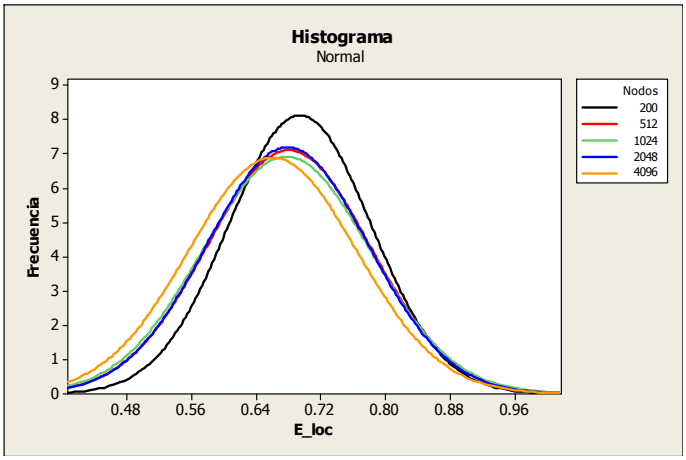


Figura b12. Histograma de la función E_{loc} por cantidad de nodos.

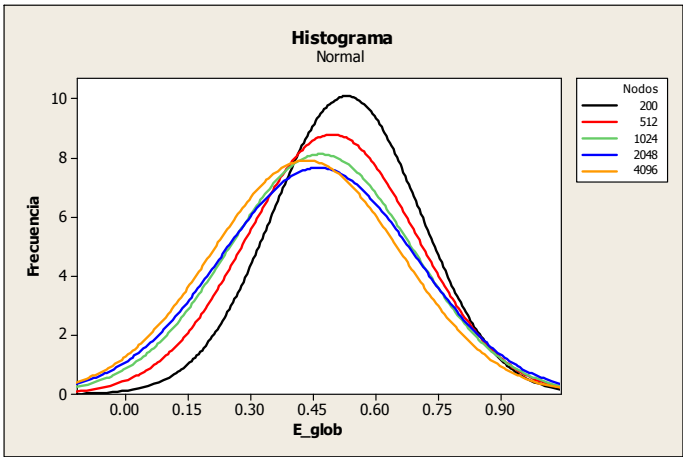


Figura b13. Histograma de la función E_{glob} por cantidad de nodos.

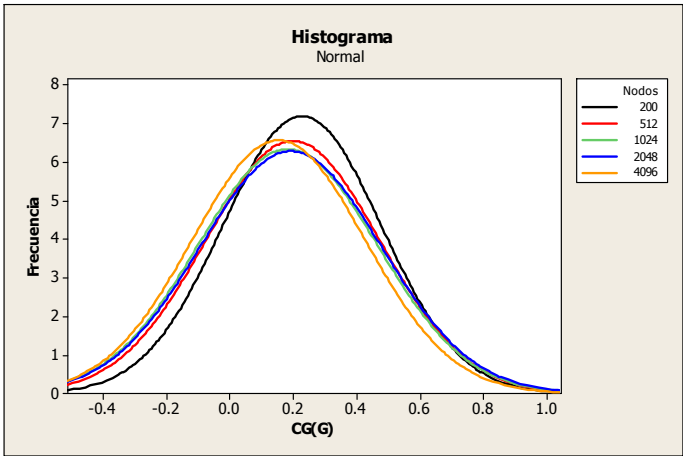


Figura b14. Histograma de la función *CG*(*G*) por cantidad de nodos.

ANEXO C

En este anexo se muestran los resultados obtenidos mediante el Análisis Multivariado de la Varianza (MANOVA) realizado en Minitab 14, de las funciones de caracterización, tanto para el modelo de los efectos fijos como para el modelo mixto.

C1. Resultados del diseño factorial modelo de los efectos fijos.

En el Cuadro c1 se presentan los resultados MANOVA para el modelo de los efectos fijos, en la primera parte se puede observar los resultados ANOVA por cada función de caracterización.

En los resultados ANOVA se observa los grados de libertad de cada elemento del modelo de los efectos fijos (DF), la suma de los cuadrados (SS) es la proporción de variación explicada por los factores, cuadrados medios (MS), el valor F_0 calculado para probar los términos en el modelo que son significativos y el valor de *P* que representa la probabilidad de cometer el error Tipo I, si el valor de *P* es pequeño significa es la probabilidad de cometer un error por rechazar la hipótesis nula es pequeña.

Después podemos observar la desviación estándar del error (S), seguida del coeficiente de determinación (R-Sq) que indica que tanta variación en el modelo es explicada por la variable respuesta y R-Sq(adj) es el coeficiente de determinación ajustado y determina que tan buena es la variable como variable predoctora, en el modelo de regresión suele ser de mayor utilidad, pero en nuestro caso también lo es.

Después se muestran las pruebas MANOVA, el estadístico Wilks, es la razón de la intervarianza de subgrupos entre la varianza total de grupo, un valor próximo a 1 significa que la intervarianza es pequeña y, por tanto, los subgrupos pueden considerarse como uno solo, un valor próximo a cero indica que la intervarianza es grande y estadísticamente los subgrupos pueden considerarse diferentes.

Para las pruebas de Hotelling-Lawley y de Roy valores altos significan diferencias significativas entre grupos, para interpretar mejor la prueba de Pillai es aconsejable observar el valor de *P* que acompaña a estas pruebas y el valor de F_0 aproximado, si estos valores son significativos, entonces las conclusiones realizadas mediante el ANOVA a cada variable se consideran correctas.

Cuadro c1. Resultados MANOVA modelo de los efectos fijos

General Linear Model: <k>, σ_<k>, DDC(G), L(G), D(G), E_loc, E_glob, CG(G) versus Tipo, Nodos

```
Factor  Type   Levels  Values
Tipo    fixed       3  1, 2, 3
Nodos   fixed       5  200,  512, 1024, 2048, 4096

Analysis of Variance for <k>
Source       DF         SS        MS       F      P
Tipo          2   56986399  28493199  218.83  0.000
Nodos         4   20324298   5081074   39.02  0.000
Tipo*Nodos    8   40552421   5069053   38.93  0.000
Error       435   56640931    130209
Total       449  174504048

S = 360.845   R-Sq = 67.54%   R-Sq(adj) = 66.50%


Analysis of Variance for σ_<k>
Source       DF        SS       MS       F      P
Tipo          2   6197.44  3098.72  327.73  0.000
Nodos         4   3529.75   882.44   93.33  0.000
Tipo*Nodos    8   3533.04   441.63   46.71  0.000
Error       435   4113.03     9.46
Total       449  17373.26

S = 3.07494   R-Sq = 76.33%   R-Sq(adj) = 75.56%


Analysis of Variance for DDC(G)
Source       DF       SS       MS         F      P
Tipo          2  53.3446  26.6723  18794.05  0.000
Nodos         4   0.5132   0.1283     90.40  0.000
Tipo*Nodos    8   1.4591   0.1824    128.51  0.000
Error       435   0.6173   0.0014
Total       449  55.9342

S = 0.0376721   R-Sq = 98.90%   R-Sq(adj) = 98.86%


Analysis of Variance for L(G)
Source       DF       SS       MS        F      P
Tipo          2  309.830  154.915  1348.78  0.000
Nodos         4   35.478    8.869    77.22  0.000
Tipo*Nodos    8   20.655    2.582    22.48  0.000
Error       435   49.962    0.115
Total       449  415.925

S = 0.338904   R-Sq = 87.99%   R-Sq(adj) = 87.60%
```

Continuación Cuadro c1. Resultados MANOVA modelo de los efectos fijos

```
Analysis of Variance for D(G)
Source       DF        SS       MS        F      P
Tipo          2   945.213  472.607  1014.56  0.000
Nodos         4    59.969   14.992    32.18  0.000
Tipo*Nodos    8    44.364    5.546    11.90  0.000
Error       435   202.633    0.466
Total       449  1252.180

S = 0.682513   R-Sq = 83.82%   R-Sq(adj) = 83.30%


Analysis of Variance for E_loc
Source       DF       SS       MS      F      P
Tipo          2  1.33479  0.66739  95.73  0.000
Nodos         4  0.05223  0.01306   1.87  0.114
Tipo*Nodos    8  0.03038  0.00380   0.54  0.823
Error       435  3.03257  0.00697
Total       449  4.44997

S = 0.0834951   R-Sq = 31.85%   R-Sq(adj) = 29.66%


Analysis of Variance for E_glob
Source       DF       SS      MS        F      P
Tipo          2  16.6855  8.3428  1066.68  0.000
Nodos         4   0.4879  0.1220    15.59  0.000
Tipo*Nodos    8   0.2724  0.0341     4.35  0.000
Error       435   3.4023  0.0078
Total       449  20.8481

S = 0.0884380   R-Sq = 83.68%   R-Sq(adj) = 83.16%


Analysis of Variance for CG(G)
Source       DF       SS       MS       F      P
Tipo          2  21.4700  10.7350  406.02  0.000
Nodos         4   0.2547   0.0637    2.41  0.049
Tipo*Nodos    8   0.1995   0.0249    0.94  0.480
Error       435  11.5012   0.0264
Total       449  33.4254

S = 0.162603   R-Sq = 65.59%   R-Sq(adj) = 64.48%


MANOVA for Tipo
s = 2    m = 2.5    n = 213.0
                        Test                DF
Criterion          Statistic          F  Num  Denom      P
Wilks'               0.00025  3355.989   16    856  0.000
Lawley-Hotelling   149.21765  3982.246   16    854  0.000
Pillai's             1.96277  2826.861   16    858  0.000
Roy's              115.29470

MANOVA for Nodos
s = 4    m = 1.5    n = 213.0
                        Test                 DF
Criterion          Statistic  Approx F  Num  Denom      P
Wilks'               0.03453    73.621   32   1579  0.000
Lawley-Hotelling    16.49926   219.904   32   1706  0.000
Pillai's             1.38144    28.422   32   1724  0.000
Roy's               15.81347
```

Continuación Cuadro c1. Resultados MANOVA modelo de los efectos fijos

```
MANOVA for Tipo*Nodos
s = 8    m = -0.5    n = 213.0
                        Test                DF
Criterion          Statistic  Approx F  Num  Denom      P
Wilks'               0.00913    48.630   64   2475  0.000
Lawley-Hotelling    15.58814   103.819   64   3410  0.000
Pillai's             2.18024    20.370   64   3480  0.000
Roy's               10.04556
```

C2. Resultados del diseño factorial modelo mixto

En el Cuadro c2 se presentan los resultados MANOVA para el modelo mixto, en la primera parte se puede observar los resultados ANOVA por cada función de caracterización.

Para el modelo mixto además de los resultados ANOVA descritos en la sección C1, se añaden las estimaciones de los componentes de la varianza para el factor aleatorio, la interacción y el error, entre mayor sea el valor del componente, mayor es la variabilidad que introduce el factor aleatorio a la función de caracterización, los valores negativos se asumirán como ceros. Las pruebas MANOVA también fueron descritas en la sección C1.

Cuadro c2. Resultados MANOVA modelo mixto

ANOVA: <k>, σ_<k>, DDC(G), L(G), D(G), E_loc, E_glob, CG(G) versus Tipo, Nodos

```
Factor  Type    Levels  Values
Tipo    fixed        3  1, 2, 3
Nodos   random       5  200,  512, 1024, 2048, 4096

Analysis of Variance for <k>
Source       DF         SS        MS      F      P
Tipo          2   56986399  28493199   5.62  0.030
Nodos         4   20324298   5081074   1.00  0.460
Tipo*Nodos    8   40552421   5069053  38.93  0.000
Error       435   56640931    130209
Total       449  174504048

S = 360.845   R-Sq = 67.54%   R-Sq(adj) = 66.50%

                                  Expected Mean Square
                Variance  Error   for Each Term (using
   Source      component   term   unrestricted model)
1  Tipo                        3  (4) + 30 (3) + Q[1]
2  Nodos             134       3  (4) + 30 (3) + 90 (2)
3  Tipo*Nodos     164628       4  (4) + 30 (3)
4  Error          130209          (4)
```

Continuación Cuadro c2. Resultados MANOVA modelo mixto

```
Analysis of Variance for σ_<k>
Source       DF        SS        MS      F      P
Tipo          2   6197.44  3098.72   7.02  0.017
Nodos         4   3529.75   882.44   2.00  0.188
Tipo*Nodos    8   3533.04   441.63  46.71  0.000
Error       435   4113.03     9.46
Total       449  17373.26

S = 3.07494   R-Sq = 76.33%   R-Sq(adj) = 75.56%

                               Expected Mean Square
                Variance  Error  for Each Term (using
   Source      component   term  unrestricted model)
1  Tipo                       3  (4) + 30 (3) + Q[1]
2  Nodos           4.898      3  (4) + 30 (3) + 90 (2)
3  Tipo*Nodos     14.406      4  (4) + 30 (3)
4  Error           9.455         (4)

Analysis of Variance for DDC(G)
Source       DF       SS       MS       F      P
Tipo          2  53.3446  26.6723  146.24  0.000
Nodos         4   0.5132   0.1283    0.70  0.611
Tipo*Nodos    8   1.4591   0.1824  128.51  0.000
Error       435   0.6173   0.0014
Total       449  55.9342

S = 0.0376721   R-Sq = 98.90%   R-Sq(adj) = 98.86%

                               Expected Mean Square
                Variance  Error  for Each Term (using
   Source      component   term  unrestricted model)
1  Tipo                       3  (4) + 30 (3) + Q[1]
2  Nodos        -0.00060      3  (4) + 30 (3) + 90 (2)
3  Tipo*Nodos    0.00603      4  (4) + 30 (3)
4  Error         0.00142         (4)

Analysis of Variance for L(G)
Source       DF       SS       MS      F      P
Tipo          2  309.830  154.915  60.00  0.000
Nodos         4   35.478    8.869   3.44  0.065
Tipo*Nodos    8   20.655    2.582  22.48  0.000
Error       435   49.962    0.115
Total       449  415.925

S = 0.338904   R-Sq = 87.99%   R-Sq(adj) = 87.60%

                               Expected Mean Square
                Variance  Error  for Each Term (using
3.44   Source      component   term  unrestricted model)
1  Tipo                       3  (4) + 30 (3) + Q[1]
2  Nodos         0.06986      3  (4) + 30 (3) + 90 (2)
3  Tipo*Nodos    0.08223      4  (4) + 30 (3)
4  Error         0.11486         (4)

Analysis of Variance for D(G)
Source       DF        SS       MS      F      P
Tipo          2   945.213  472.607  85.22  0.000
Nodos         4    59.969   14.992   2.70  0.108
Tipo*Nodos    8    44.364    5.546  11.90  0.000
Error       435   202.633    0.466
Total       449  1252.180
```

Continuación Cuadro c2. Resultados MANOVA modelo mixto

```
S = 0.682513   R-Sq = 83.82%   R-Sq(adj) = 83.30%

                                 Expected Mean Square
                  Variance  Error  for Each Term (using
   Source        component   term  unrestricted model)
1  Tipo                         3  (4) + 30 (3) + Q[1]
2  Nodos           0.1050       3  (4) + 30 (3) + 90 (2)
3  Tipo*Nodos      0.1693       4  (4) + 30 (3)
4  Error           0.4658          (4)

Analysis of Variance for E_loc
Source       DF       SS       MS       F      P
Tipo          2  1.33479  0.66739  175.77  0.000
Nodos         4  0.05223  0.01306    3.44  0.065
Tipo*Nodos    8  0.03038  0.00380    0.54  0.823
Error       435  3.03257  0.00697
Total       449  4.44997

S = 0.0834951   R-Sq = 31.85%   R-Sq(adj) = 29.66%

                                 Expected Mean Square
                  Variance  Error  for Each Term (using
   Source        component   term  unrestricted model)
1  Tipo                         3  (4) + 30 (3) + Q[1]
2  Nodos           0.00010      3  (4) + 30 (3) + 90 (2)
3  Tipo*Nodos     -0.00011      4  (4) + 30 (3)
4  Error           0.00697         (4)

Analysis of Variance for E_glob
Source       DF       SS      MS       F      P
Tipo          2  16.6855  8.3428  244.98  0.000
Nodos         4   0.4879  0.1220    3.58  0.059
Tipo*Nodos    8   0.2724  0.0341    4.35  0.000
Error       435   3.4023  0.0078
Total       449  20.8481

S = 0.0884380   R-Sq = 83.68%   R-Sq(adj) = 83.16%

                                 Expected Mean Square
                  Variance  Error  for Each Term (using
   Source        component   term  unrestricted model)
1  Tipo                         3  (4) + 30 (3) + Q[1]
2  Nodos           0.00098      3  (4) + 30 (3) + 90 (2)
3  Tipo*Nodos      0.00087      4  (4) + 30 (3)
4  Error           0.00782         (4)

Analysis of Variance for CG(G)
Source       DF       SS       MS       F      P
Tipo          2  21.4700  10.7350  430.39  0.000
Nodos         4   0.2547   0.0637    2.55  0.121
Tipo*Nodos    8   0.1995   0.0249    0.94  0.480
Error       435  11.5012   0.0264
Total       449  33.4254

S = 0.162603   R-Sq = 65.59%   R-Sq(adj) = 64.48%

                                 Expected Mean Square
                  Variance  Error  for Each Term (using
   Source        component   term  unrestricted model)
1  Tipo                         3  (4) + 30 (3) + Q[1]
2  Nodos           0.00043      3  (4) + 30 (3) + 90 (2)
```

Continuación Cuadro c2. Resultados MANOVA modelo mixto

```
3  Tipo*Nodos   -0.00005      4  (4) + 30 (3)
4  Error         0.02644         (4)


MANOVA for Tipo
s = 2    m = 2.5    n = -0.5
                                              DF
Criterion         Test Statistic        F  Num  Denom      P
Wilks'                   0.00000  180.977   16      2  0.006
Lawley-Hotelling      5630.48365    0.000   16      0      *
Pillai's                 1.99732  186.086   16      4  0.000
Roy's                 5230.22719

These tests use error term = Tipo*Nodos


MANOVA for Nodos
s = 4    m = 1.5    n = -0.5

                      Test                 DF
Criterion         Statistic  Approx F  Num  Denom      P
Wilks'              0.00000     5.947   32      5  0.028
Lawley-Hotelling  564.18965    -8.815   32     -2      *
Pillai's            3.41144     2.898   32     16  0.014
Roy's             481.94557

These tests use error term = Tipo*Nodos


MANOVA for Tipo*Nodos
s = 8    m = -0.5    n = 213.0

                      Test                 DF
Criterion         Statistic  Approx F  Num  Denom      P
Wilks'              0.00913    48.630   64   2475  0.000
Lawley-Hotelling   15.58814   103.819   64   3410  0.000
Pillai's            2.18024    20.370   64   3480  0.000
Roy's              10.04556
```

ANEXO D

Como ayuda para la interpretación practica de cualquier experimento donde intervengan dos o más factores, es conveniente construir gráficas de los efectos principales y la interacción entre los factores. En este anexo se muestran las gráficas de los efectos principales y de la interacción para cada una de las funciones de caracterización, el software utilizado para obtener estas gráficas fue MINITAB 14.

En las gráficas de los efectos principales es importante observar la dirección (positiva o negativa) y la magnitud de los efectos. Los efectos principales serán positivos si al aumentar un nivel del factor la variable respuesta aumenta, o negativos si al aumentar un nivel de factor la variable respuesta disminuye. La magnitud de los efectos se observa comparando entre si estas gráficas.

En las gráficas de la interacción entre los factores, es de interés observar si existe una interacción entre los factores tal que influyan en los valores de las variables respuesta, si las líneas son paralelas, se dice que no hay interacción entre los factores y bastará con interpretar solo las gráficas de los efectos principales, cuando las líneas se cruzan se dice que si hay interacción.

D1. Gráficas de los efectos principales y de la interacción de factores

La Figura d1, muestra la gráfica de los efectos principales para la función $\langle k \rangle$, donde se puede observar que tanto el factor tipo de red como el factor número de nodos tienen grandes efectos sobre esta función de caracterización, se puede concluir que para las redes Exponenciales (3) y las Power-Law (2) la media es menor en comparación con las redes Aleatorias (1) y que conforme se incrementa el número de nodos la media de la función $\langle k \rangle$ aumenta.

La Figura d2, muestra la gráfica de interacción de los factores, donde se puede observar que en las redes Exponenciales (3) y Power-Law (2) no existe interacción, pero en las redes Aleatorias (1) se aprecia que el número de nodos si afecta significativamente a esta función.

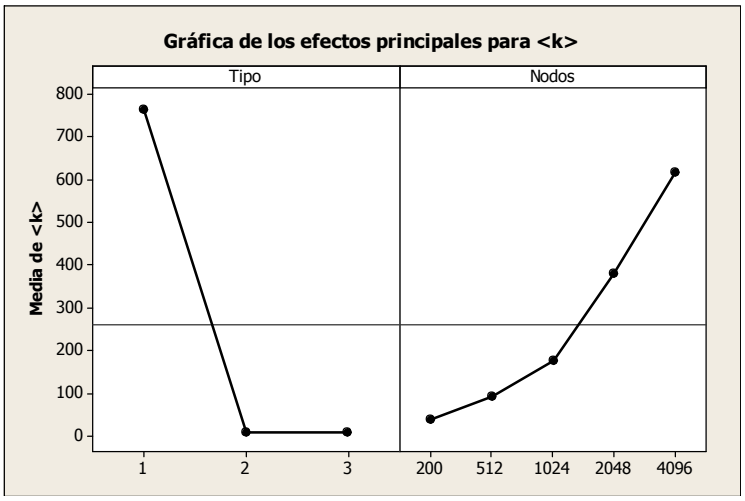


Figura d1. Gráfica de los efectos principales para la función $\langle k \rangle$

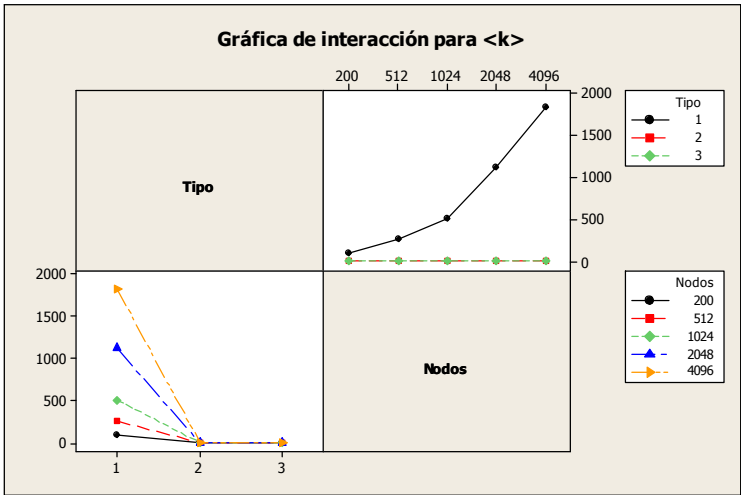


Figura d2. Gráfica de la interacción de factores para la función $\langle k \rangle$

La Figura d3, muestra la gráfica de los efectos principales para la función $\sigma_{\langle k \rangle}$, tanto el factor tipo de red como el factor número de nodos tienen grandes efectos en esta función, se puede concluir que conforme aumenta el número de nodos la media de la función $\sigma_{\langle k \rangle}$

aumenta. En la Figura d4 se muestra la gráfica de la interacción donde se puede apreciar que hay cierta interacción entre el tipo de red y el número de nodos.

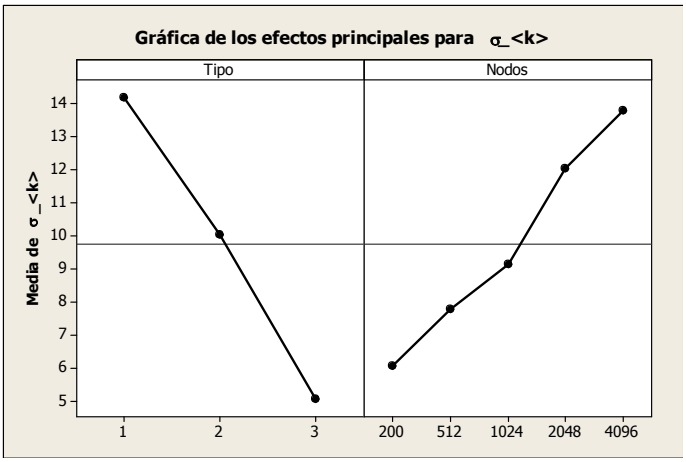


Figura d3. Gráfica de los efectos principales para la función $\sigma_{\langle k \rangle}$

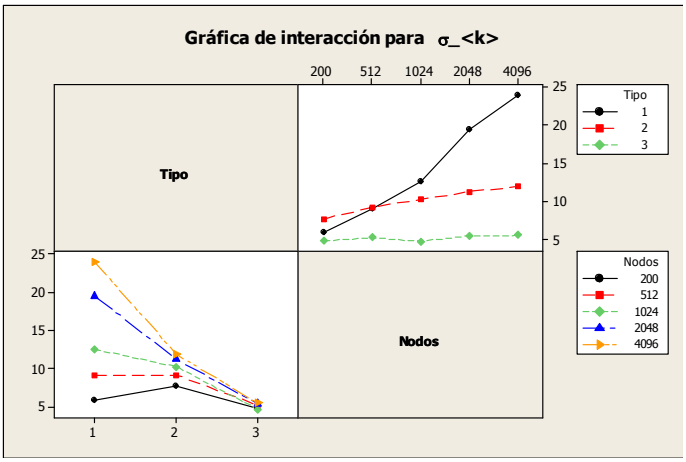


Figura d4. Gráfica de la interacción de factores para la función $\sigma_{\langle k \rangle}$

En las Figuras d5 y d7 se muestran las gráficas de los efectos principales para las funciones $L(G)$ y $D(G)$ respectivamente, en estas funciones tanto el factor tipo de red como el factor número de nodos tienen grandes efectos. En las Figuras d6 y d8 se muestran las gráficas de la interacción de factores para las funciones $L(G)$ y $D(G)$ respectivamente, en estas funciones se puede observar cierta interacción entre el factor tipo de red y el factor número de nodos en las redes Exponenciales (3) y Power-Law (2).

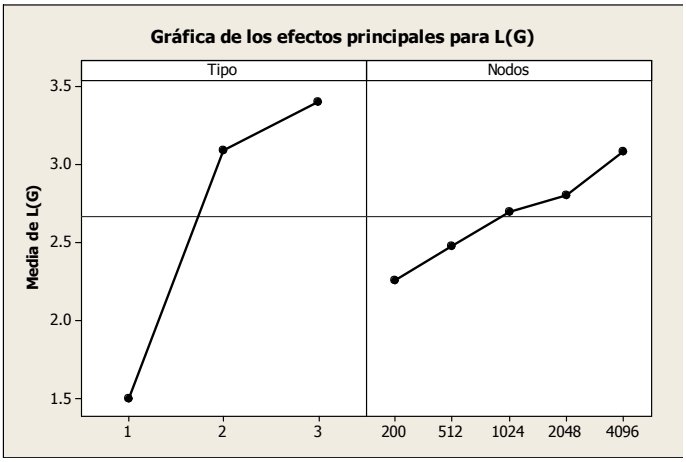


Figura d5. Gráfica de los efectos principales para la función *L*(*G*)

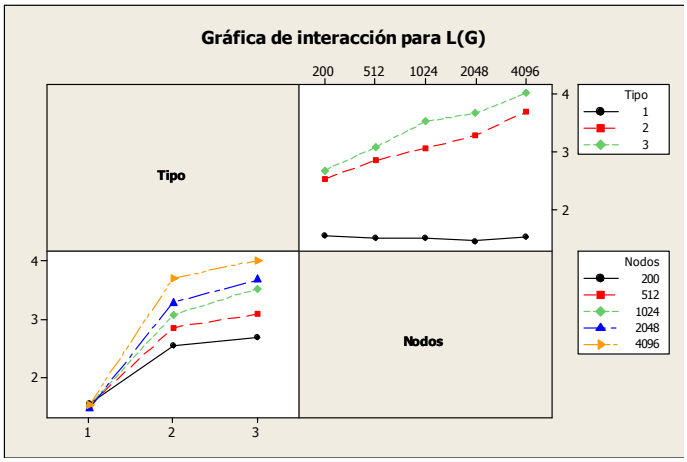


Figura d6. Gráfica de la interacción de factores para la función *L*(*G*)

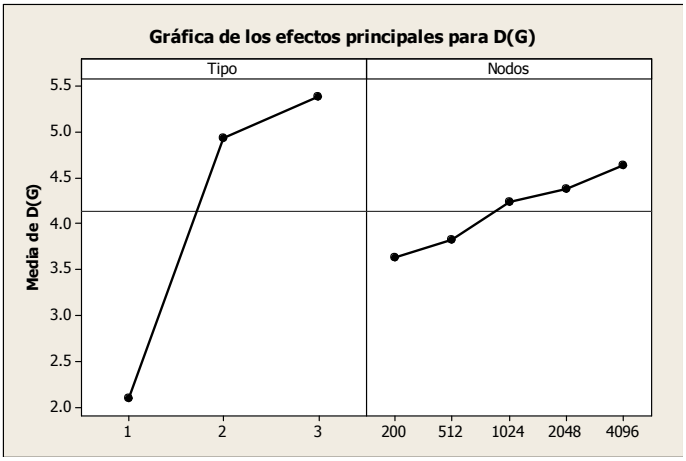


Figura d7. Gráfica de los efectos principales para la función *D*(*G*)

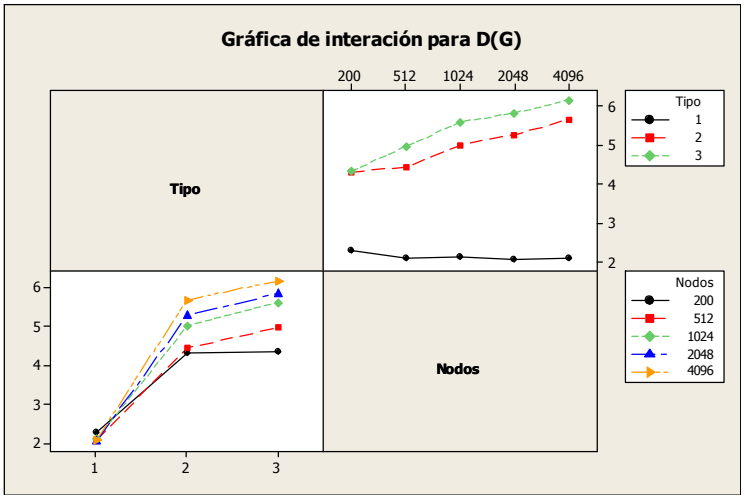


Figura d8. Gráfica de la interacción de factores para la función *D*(*G*)

En las Figuras d9, d11 y d13 se muestran las gráficas de los efectos principales para las funciones E_{loc}, E_{glob} y $CG(G)$ respectivamente, en estas funciones el tipo de red tiene mucho mayor que el número de nodos, a medida que aumenta el número de nodos estas funciones se ven afectadas ligeramente.

En las Figuras d10, d12 y d14 se muestran las gráficas de interacción de factores para las funciones E_{loc}, E_{glob} y $CG(G)$ respectivamente, en estas funciones se puede observar que existe cierta interacción entre el tipo de red y el número de nodos.

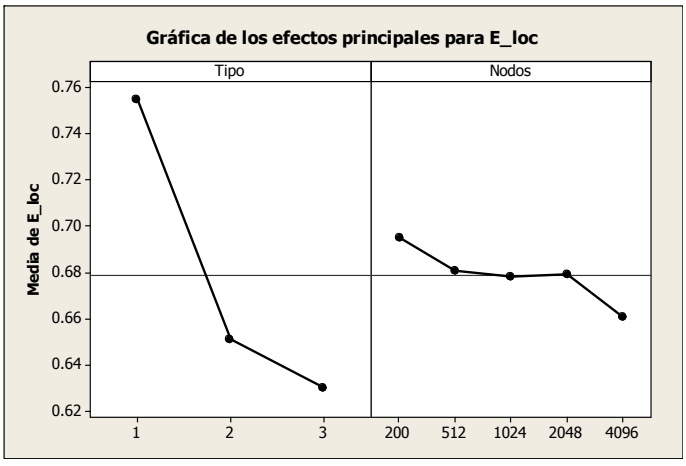


Figura d9. Gráfica de los efectos principales para la función E_{loc}

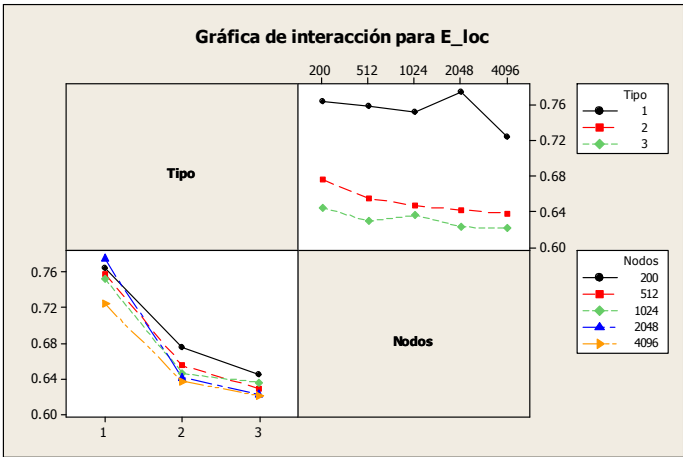


Figura d10. Gráfica de la interacción de factores para la función E_{loc}

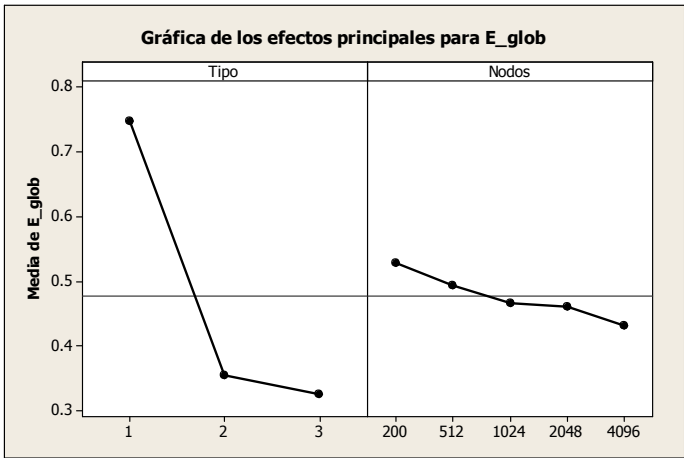


Figura d11. Gráfica de los efectos principales para la función E_{glob}

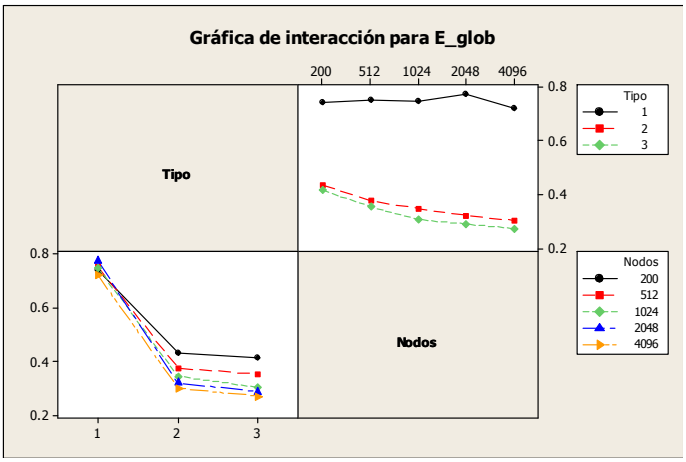


Figura d12. Gráfica de la interacción de factores para la función E_{glob}

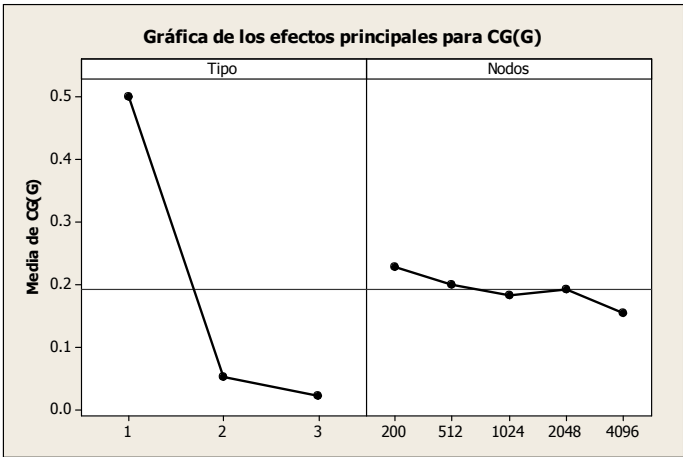


Figura d13. Gráfica de los efectos principales para la función *CG*(*G*)

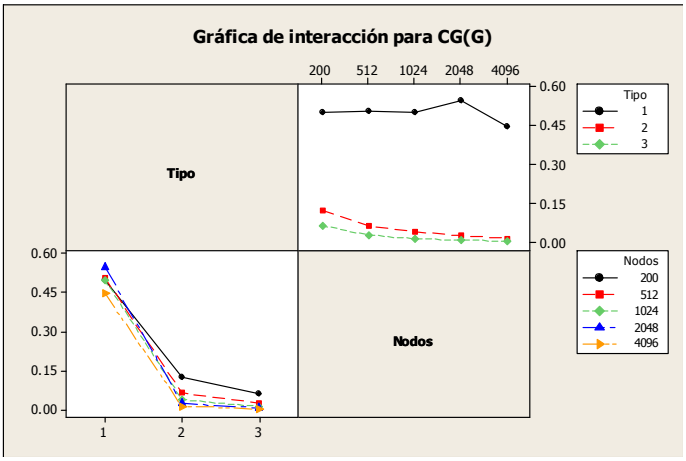


Figura d14. Gráfica de la interacción de factores para la función *CG*(*G*)

ANEXO E

En este anexo se muestran las gráficas de los residuales para cada una de las funciones de caracterización, el software utilizando para obtener estas gráficas fue MINITAB 14. El análisis de los residuales permite descubrir inadecuaciones del modelo y violaciones a los supuestos de normalidad e independencia.

Para verificar el supuesto de normalidad se gráfica un histograma de los residuales, si se satisface este supuesto esta gráfica debe aparecer como una muestra de la distribución normal con centro cero. También es útil construir una gráfica de probabilidad normal de los residuales, la cual deberá tener la apariencia de una línea recta, para visualizar la línea recta se deberá prestar más atención a los valores centrales de la gráfica que a los valores extremos [Montgomery 2004].

Para verificar el supuesto de independencia los residuales gráficados en contra de sus valores ajustados y el orden en que se recolectaron deberán estar sin estructura, es decir no deberán contener patrones obvios, también se debe verificar que no estén relacionados con ninguna otra variable, incluyendo a los factores.

E1. Gráficas de los residuales para cada función de caracterización

A continuación se presentan las gráficas de la probabilidad normal de los residuales junto con su histograma, los residuales graficados contra los valores ajustados, el orden de la observación, el tipo y número de nodos.

En la Figura e1, se observa la gráfica de los residuales para la función $\langle k \rangle$, donde se puede apreciar que los residuales no se distribuyen normalmente y se observa un patrón en los datos por lo que se puede decir que esta función no satisface los supuestos de normalidad e independencia, de esta manera las conclusiones realizadas en la sección 7.2.1.3 para esta función no serán tomadas en cuenta.

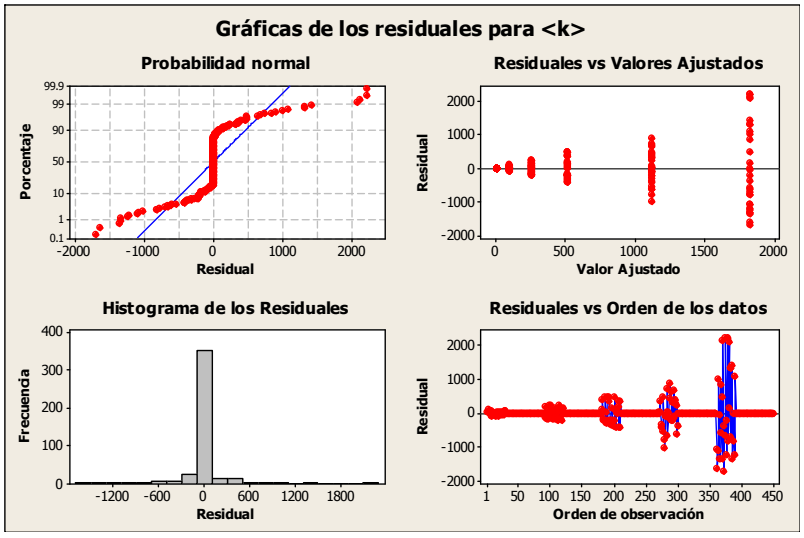
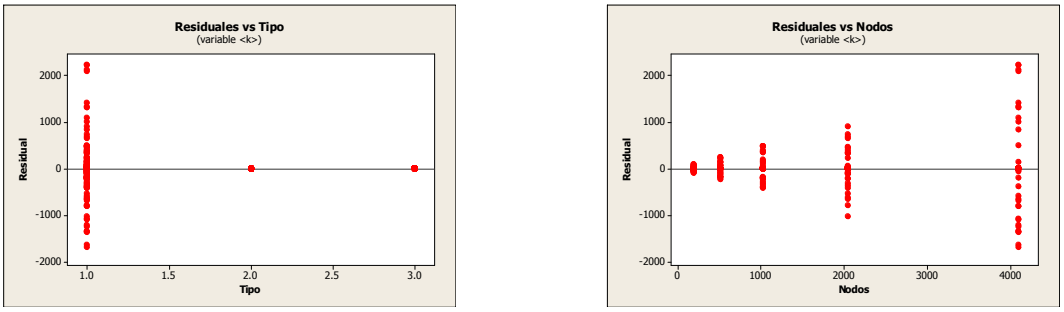


Figura e1. Gráfica de los residuales para la función $\langle k \rangle$

En la Figura e2*a*) se muestra la gráfica de los residuales para la función $\langle k \rangle$ por tipo de red y en la Figura e2*b*) se muestra la gráfica de los residuales para la función $\langle k \rangle$ por número de nodos, en estas gráficas se puede observar que existe un patrón en los datos, por lo que se puede decir que los datos obtenidos mediante la función $\langle k \rangle$ no son independientes del tipo de red y del número de nodos existentes en la red.



a) *b*)

Figura e2. Gráfica de los residuales para la función $\langle k \rangle$ *a*) por tipo de red, *b*) por número de nodos

En la Figura e3, se observa la gráfica de los residuales para la función $\sigma_{\langle k \rangle}$, donde se puede apreciar que los residuales se distribuyen normalmente y no se observa un patrón obvio en los datos por lo que se puede decir que esta función satisface los supuestos de

normalidad e independencia, de esta manera las conclusiones realizadas en la sección 7.2.1.3 son correctas.

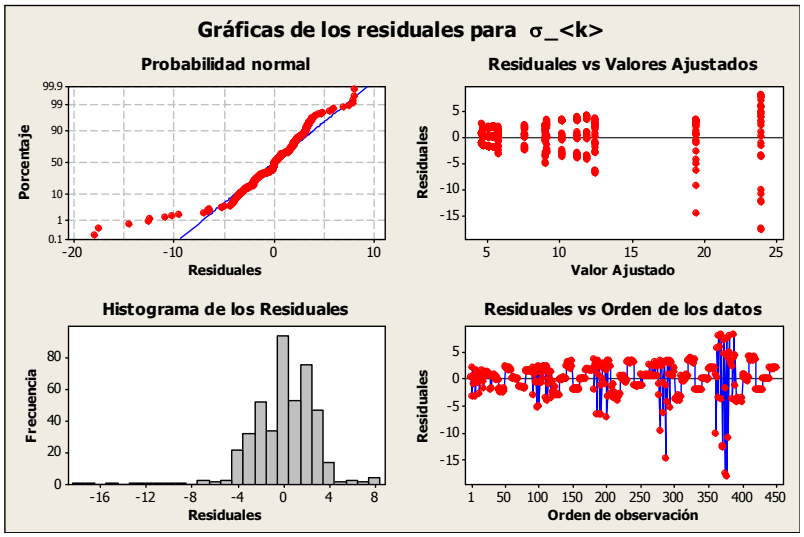
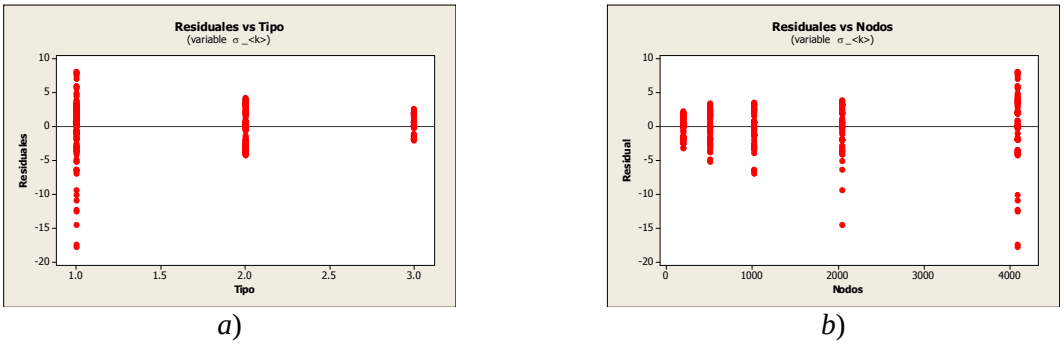


Figura e3. Gráfica de los residuales para la función $\sigma_{\langle k \rangle}$

En la Figura e4*a*) se muestra la gráfica de los residuales para la función $\sigma_{\langle k \rangle}$ por tipo de red y en la Figura e4*b*) se muestra la gráfica de los residuales para la función $\sigma_{\langle k \rangle}$ por número de nodos, en estas gráficas se puede observar que no existe ningún patrón obvio, por lo que se puede decir que los datos obtenidos mediante la función $\sigma_{\langle k \rangle}$ son independientes del tipo de red y del número de nodos existentes en la red.



a) *b*)

Figura e4. Gráfica de los residuales para la función $\sigma_{\langle k \rangle}$ *a*) por tipo de red, *b*) por número de nodos

En la Figura e6, se observa la gráfica de los residuales para la función $L(G)$, donde se puede apreciar que los residuales se distribuyen normalmente y no se observa un patrón obvio en los datos por lo que se puede decir que esta función satisface los supuestos de normalidad e independencia, de esta manera las conclusiones realizadas en la sección 7.2.1.3 son correctas.

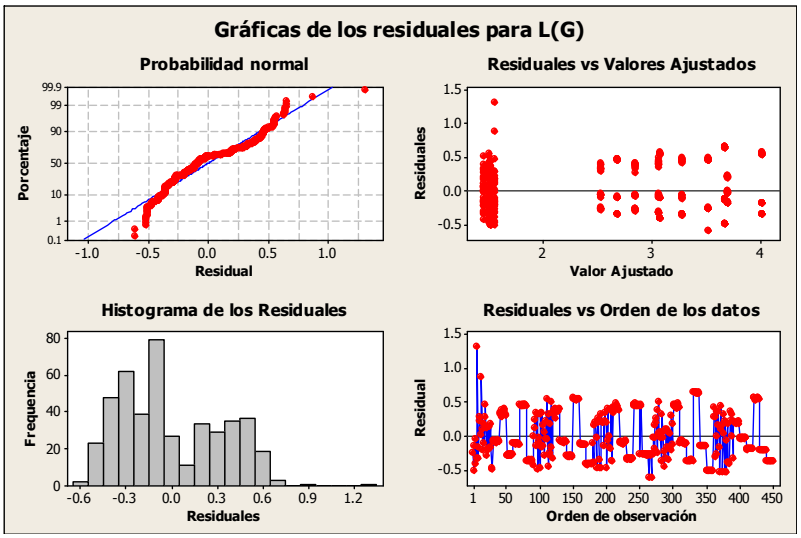
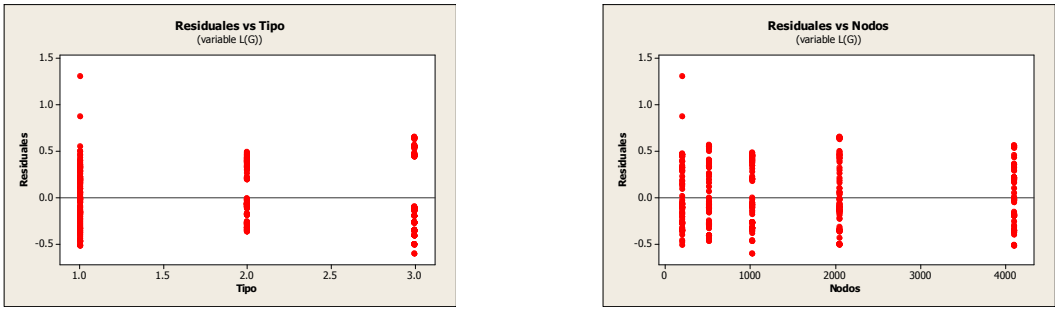


Figura e5. Gráfica de los residuales para la función $L(G)$



a) *b*)

Figura e6. Gráfica de los residuales para la función $L(G)$ *a*) por tipo de red, *b*) por número de nodos

En la Figura e6*a*) se muestra la gráfica de los residuales para la función $L(G)$ por tipo de red y en la Figura e6*b*) se muestra la gráfica de los residuales para la función $L(G)$ por número de nodos, en estas gráficas se puede observar que no existe ningún patrón

obvio, por lo que se puede decir que los datos obtenidos mediante la función $L(G)$ son independientes del tipo de red y del número de nodos existentes en la red.

En la Figura e7, se observa la gráfica de los residuales para la función $D(G)$, donde se puede apreciar que los residuales se distribuyen normalmente y no se observa un patrón obvio en los datos por lo que se puede decir que esta función satisface los supuestos de normalidad e independencia, de esta manera las conclusiones realizadas en la sección 7.2.1.3 son correctas.

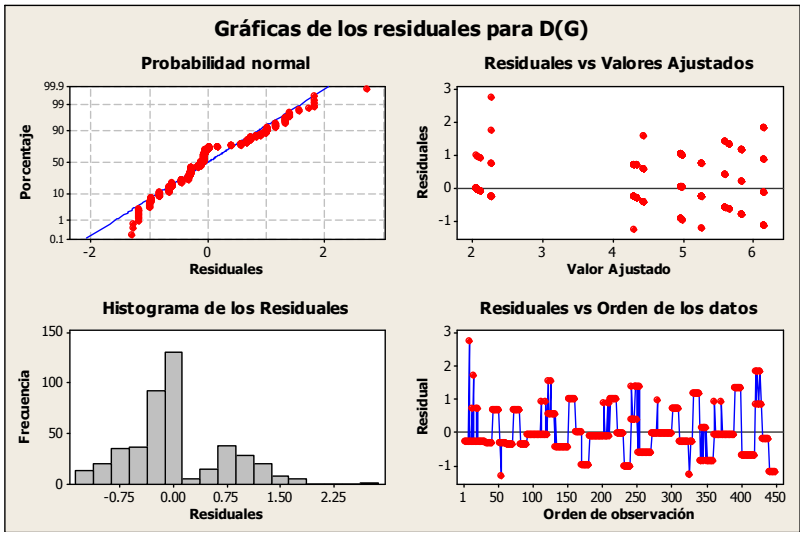
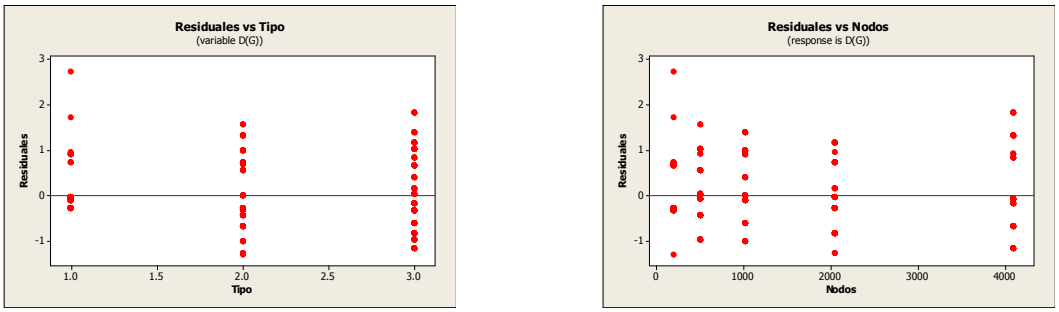


Figura e7. Gráfica de los residuales para la función *D*(*G*)



a) *b*)

Figura e8. Gráfica de los residuales para la función *D*(*G*) *a*) por tipo de red, *b*) por número de nodos

En la Figura e8*a*) se muestra la gráfica de los residuales para la función $D(G)$ por tipo de red y en la Figura e8*b*) se muestra la gráfica de los residuales para la función $D(G)$ por número de nodos, en estas gráficas se puede observar que no existe ningún patrón obvio, por lo que se puede decir que los datos obtenidos mediante la función $D(G)$ son independientes del tipo de red y del número de nodos existentes en la red.

En la Figura e9, se observa la gráfica de los residuales para la función E_{loc}, donde se puede apreciar que los residuales se distribuyen normalmente y no se observa un patrón obvio en los datos por lo que se puede decir que esta función satisface los supuestos de normalidad e independencia, de esta manera las conclusiones realizadas en la sección 7.2.1.3 son correctas.

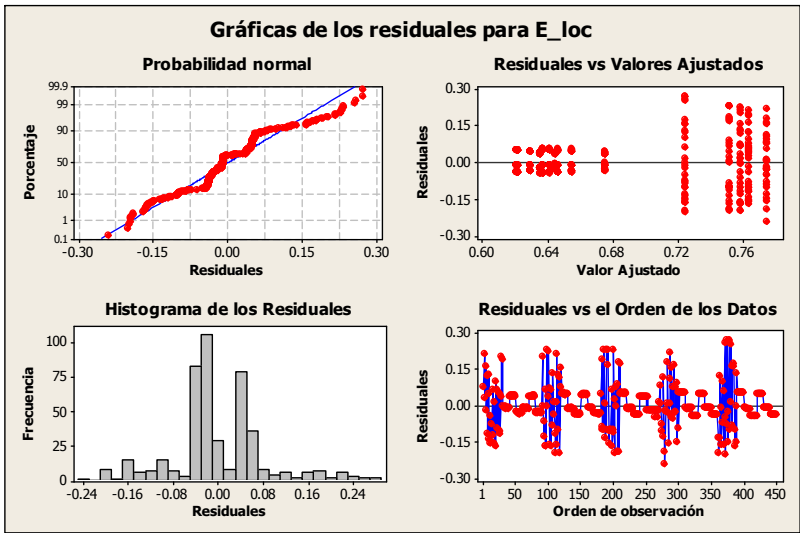
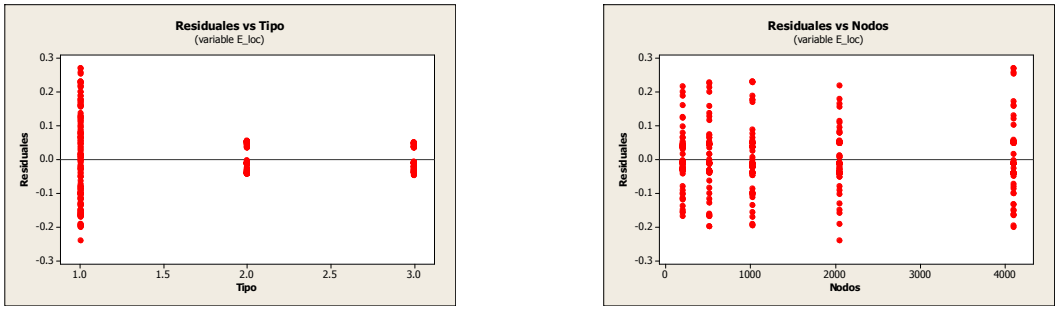


Figura e9. Gráfica de los residuales para la función E_{loc}

En la Figura e10a) se muestra la gráfica de los residuales para la función E_{loc} por tipo de red y en la Figura e10b) se muestra la gráfica de los residuales para la función E_{loc} por número de nodos, en estas gráficas se puede observar que no existe ningún patrón obvio, por lo que se puede decir que los datos obtenidos mediante la función E_{loc} son independientes del tipo de red y del número de nodos existentes en la red.



a) *b*)

Figura e10. Gráfica de los residuales para la función E_{loc} *a*) por tipo de red, *b*) por número de nodos

En la Figura e11, se observa la gráfica de los residuales para la función E_{glob}, donde se puede apreciar que los residuales se distribuyen normalmente y no se observa un patrón obvio en los datos por lo que se puede decir que esta función satisface los supuestos de normalidad e independencia, de esta manera las conclusiones realizadas en la sección 7.2.1.3 son correctas.

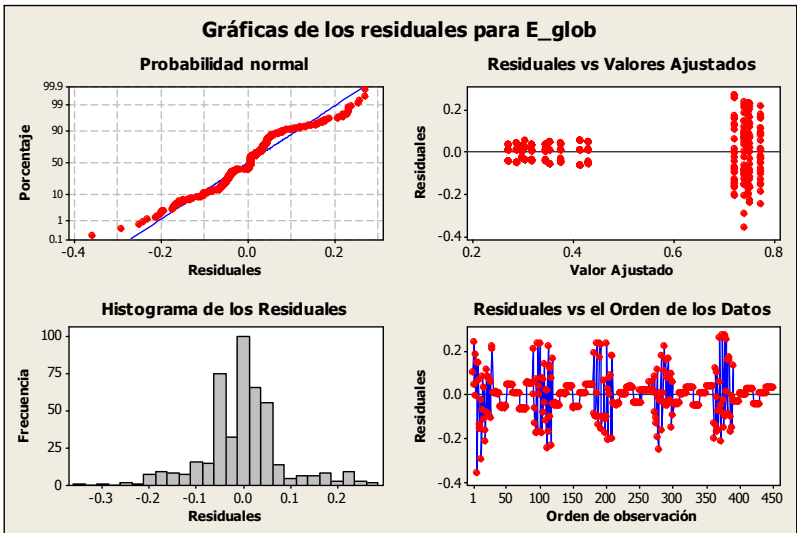
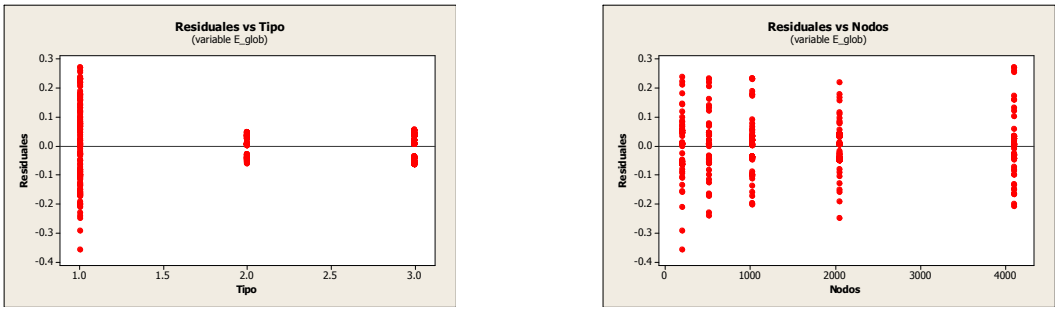


Figura e11. Gráfica de los residuales para la función E_{glob}

En la Figura e12a) se muestra la gráfica de los residuales para la función E_{glob} por tipo de red y en la Figura e12b) se muestra la gráfica de los residuales para la función E_{glob} por número de nodos, en estas gráficas se puede observar que no existe ningún patrón

obvio, por lo que se puede decir que los datos obtenidos mediante la función E_{glob} son independientes del tipo de red y del número de nodos existentes en la red.



a) *b*)

Figura e12. Gráfica de los residuales para la función E_{glob} *a*) por tipo de red, *b*) por número de nodos

En la Figura e13, se observa la gráfica de los residuales para la función *CG*(*G*), donde se puede apreciar que los residuales no se distribuyen normalmente y se observa un patrón en los datos por lo que se puede decir que esta función no satisface los supuestos de normalidad e independencia, de esta manera las conclusiones realizadas en la sección 7.2.1.3 para esta función no serán tomadas en cuenta.

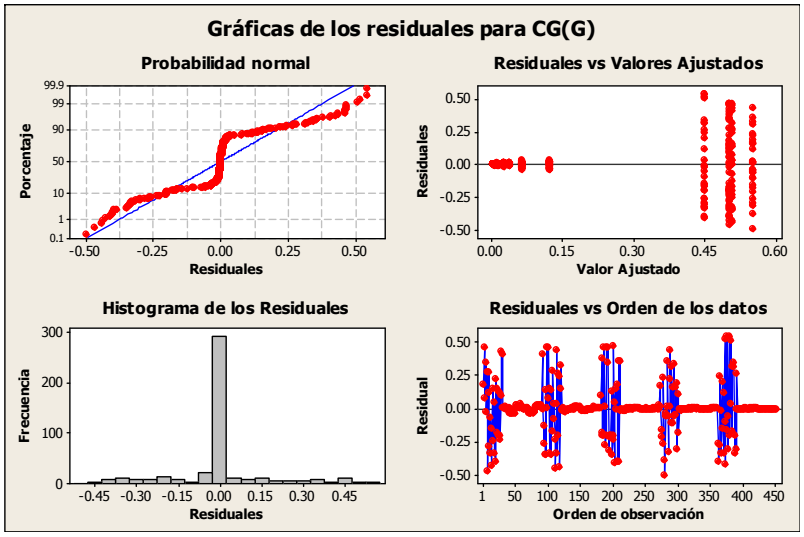
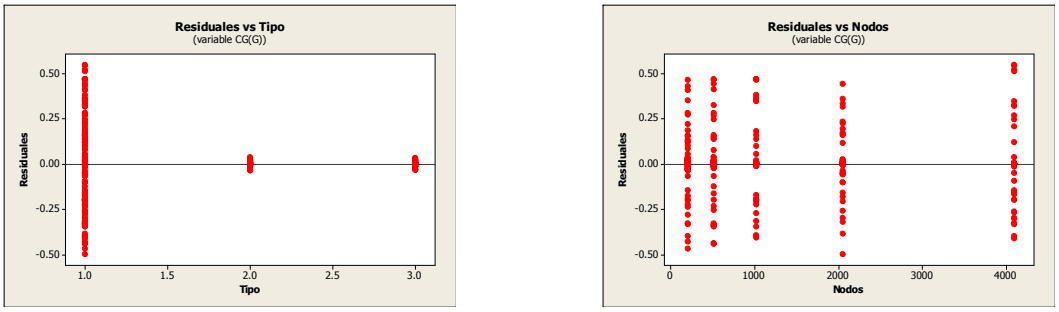


Figura e13. Gráfica de los residuales para la función *CG*(*G*)

En la Figura e14*a*) se muestra la gráfica de los residuales para la función *CG*(*G*) por tipo de red en esta gráfica se puede observar que existen un patrón en los datos, por lo que se puede decir que los datos obtenidos mediante esta función no son independientes del tipo de red. Por otra parte en la Figura e14*b*) se muestra la gráfica de los residuales para la función *CG*(*G*) por número de nodos, en esta gráfica se puede observar que no existe ningún patrón obvio, por lo que se puede decir que los datos obtenidos mediante esta función son independientes del número de nodos existentes en la red.



a) *b*)

Figura e14. Gráfica de los residuales para la función *CG*(*G*) *a*) por tipo de red, *b*) por número de nodos

Las violaciones a los supuestos de normalidad e independencias que se pueden observar en las funciones $\langle k \rangle$ y *CG*(*G*) En las Figuras d1, d2, d15 y d16 corresponden a las gráficas de los residuales para $\langle k \rangle$ y *CG*(*G*) en estas se puede observar violaciones a los supuestos de normalidad e independencia, esto porque a pesar de que no se determino el número de aristas, $\langle k \rangle$ y *CG*(*G*) están definidos por parámetros de inicio de los modelos empleados para la generación de redes Power-Law y redes Exponenciales, por lo cual las conclusiones realizadas para estas variables quedaron descartadas.

ANEXO F

Este anexo contiene los resultados obtenidos mediante las comparaciones múltiples de las medias de las funciones de caracterización seleccionadas en la sección empleando la prueba de Tukey, el software utilizando para realizar las comparaciones múltiples SAS V8.

Las comparaciones múltiples son útiles cuando se desean detectar diferencias específicas entre los diferentes niveles de los factores, esto se logra haciendo comparaciones entre los pares de medias, el procedimiento para efectuar la prueba de Tukey se encuentra detallado en [Montgomery 2004]. En el Cuadro f1 se muestran los resultados de las comparaciones múltiples efectuadas sobre las funciones de caracterización.

Cuadro f1. Resultados de la Prueba de Tukey para las funciones de caracterización

Tukey's Studentized Range (HSD) Test for σ_<k>

Type II error rate than REGWQ.

Alpha	0.01
Error Degrees of Freedom	435
Error Mean Square	9.455247
Critical Value of Studentized Range	4.14212
Minimum Significant Difference	1.04

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	Type
A	14.1708	150	1
B	10.0332	150	2
C	5.0924	150	3

Tukey's Studentized Range (HSD) Test for DDC(G)

Type II error rate than REGWQ.

Alpha	0.01
Error Degrees of Freedom	435
Error Mean Square	0.001419
Critical Value of Studentized Range	4.1421
Minimum Significant Difference	0.0127

Continuación Cuadro f1. Resultados de la Prueba de Tukey para las funciones de caracterización

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	Type
A	0.888839	150	2
B	0.445879	150	3
C	0.045840	150	1

Tukey's Studentized Range (HSD) Test for L(G)

Type II error rate than REGWQ.

Alpha	0.01
Error Degrees of Freedom	435
Error Mean Square	0.114856
Critical Value of Studentized Range	4.14212
Minimum Significant Difference	0.1146

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	Type
A	3.39874	150	3
B	3.08950	150	2
C	1.50441	150	1

Tukey's Studentized Range (HSD) Test for D(G)

Type II error rate than REGWQ.

Alpha	0.01
Error Degrees of Freedom	435
Error Mean Square	0.465824
Critical Value of Studentized Range	4.14212
Minimum Significant Difference	0.2308

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	Type
A	5.38000	150	3
B	4.93333	150	2
C	2.10667	150	1

Tukey's Studentized Range (HSD) Test for E_loc

Type II error rate than REGWQ.

Alpha	0.01
Error Degrees of Freedom	435
Error Mean Square	0.006971
Critical Value of Studentized Range	4.14212
Minimum Significant Difference	0.0282

Continuación Cuadro f1. Resultados de la Prueba de Tukey para las funciones de caracterización

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	Type
A	0.754855	150	1
B	0.651121	150	2
B			
B	0.630343	150	3

Tukey's Studentized Range (HSD) Test for E_glob

Type II error rate than REGWQ.

Alpha	0.01
Error Degrees of Freedom	435
Error Mean Square	0.007821
Critical Value of Studentized Range	4.14212
Minimum Significant Difference	0.0299

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	Type
A	0.74833	150	1
B	0.35495	150	2
B			
B	0.32627	150	3

ANEXO G

Este anexo contiene los resultados de la clasificación de redes complejas de naturaleza desconocida. El algoritmos de clasificación empleado fue el Naive Bayes, el modelo fue construido mediante Weka, la función de caracterización utilizada como entrada para el clasificador Naive Bayes fue el coeficiente de dispersión del grado $DDC(G)$.

La Tabla g1 contiene los resultados de la clasificación, en la primera columna se tiene el nombre de la instancia de red, enseguida se muestra el tipo de red que el clasificador asignó a la instancia de acuerdo al *DDC(G)* en la tercera y cuarta columna se muestra el modelo utilizado para generar la instancia y que tipo de red reproduce, y en la última columna se muestra si el resultado de la clasificación fue correcto (1) o incorrecto (0). Al final de la tabla se muestra el número de clasificaciones correctas.

Tabla g1. Resultados de la clasificación de redes complejas de naturaleza desconocida.

Clasificación de Redes Complejas de Naturaleza Desconocida				
Instancias de red	**Resultados de la Clasificación**	**Información de las redes generadas**		**Resultados de la Clasificación Correctos**
	Tipo de Red	**Modelo de generación**	**Tipo de red**	
inst_elisa_01	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_02	Redes Aleatorias	WS	Redes Aleatorias	1
inst_elisa_03	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_04	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_05	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_06	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_07	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_08	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_09	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_10	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_11	Exponencial	BA	Redes Power - Law	0
inst_elisa_12	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_13	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_14	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_15	Redes Aleatorias	Kleinberg	Redes Aleatorias	1

Continuación Tabla g1. Resultados de la clasificación de redes complejas de naturaleza desconocida.

inst_elisa_16	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_17	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_18	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_19	Redes Aleatorias	WS	Redes Aleatorias	1
inst_elisa_20	Redes Aleatorias	WS	Redes Aleatorias	1
inst_elisa_21	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_22	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_23	Redes Aleatorias	WS	Redes Aleatorias	1
inst_elisa_24	Redes Aleatorias	WS	Redes Aleatorias	1
inst_elisa_25	Redes Aleatorias	WS	Redes Aleatorias	1
inst_elisa_26	Redes Aleatorias	WS	Redes Aleatorias	1
inst_elisa_27	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_28	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_29	Redes Aleatorias	WS	Redes Aleatorias	1
inst_elisa_30	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_31	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_32	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_33	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_34	Redes Aleatorias	WS	Redes Aleatorias	1
inst_elisa_35	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_36	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_37	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_38	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_39	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_40	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_41	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_42	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_43	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_44	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_45	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_46	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_47	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_48	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_49	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_50	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_51	Redes Aleatorias	Geometrical	Redes Power - Law	0
inst_elisa_52	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_53	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_54	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_55	Redes Aleatorias	WS	Redes Aleatorias	1
inst_elisa_56	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_57	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_58	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_59	Redes Power - Law	Geometrical	Redes Power - Law	1

Continuación Tabla g1. Resultados de la clasificación de redes complejas de naturaleza desconocida.

inst_elisa_60	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_61	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_62	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_63	Redes Aleatorias	WS	Redes Aleatorias	1
inst_elisa_64	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_65	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_66	Redes Aleatorias	WS	Redes Aleatorias	1
inst_elisa_67	Redes Aleatorias	WS	Redes Aleatorias	1
inst_elisa_68	Redes Aleatorias	WS	Redes Aleatorias	1
inst_elisa_69	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_70	Redes Power - Law	Aleatorios	Redes Aleatorias	0
inst_elisa_71	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_72	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_73	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_74	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_75	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_76	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_77	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_78	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_79	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_80	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_81	Redes Aleatorias	SW	Redes Aleatorias	1
inst_elisa_82	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_83	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_84	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_85	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_86	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_87	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_88	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_89	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_90	Redes Aleatorias	SW	Redes Aleatorias	1
inst_elisa_91	Redes Aleatorias	Geometrical	Redes Power - Law	0
inst_elisa_92	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_93	Redes Aleatorias	Geometrical	Redes Power - Law	0
inst_elisa_94	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_95	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_96	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_97	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_98	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_99	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_100	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_101	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_102	Redes Aleatorias	SW	Redes Aleatorias	1
inst_elisa_103	Redes Power - Law	BA	Redes Power - Law	1

Continuación Tabla g1. Resultados de la clasificación de redes complejas de naturaleza desconocida.

inst_elisa_104	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_105	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_106	Redes Aleatorias	Geometrical	Redes Power - Law	0
inst_elisa_107	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_108	Redes Aleatorias	Geometrical	Redes Power - Law	0
inst_elisa_109	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_110	Redes Aleatorias	SW	Redes Aleatorias	1
inst_elisa_111	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_112	Redes Aleatorias	SW	Redes Aleatorias	1
inst_elisa_113	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_114	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_115	Redes Aleatorias	Geometrical	Redes Power - Law	0
inst_elisa_116	Redes Aleatorias	Geometrical	Redes Power - Law	0
inst_elisa_117	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_118	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_119	Redes Aleatorias	SW	Redes Aleatorias	1
inst_elisa_120	Redes Aleatorias	SW	Redes Aleatorias	1
inst_elisa_121	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_122	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_123	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_124	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_125	Redes Aleatorias	Geometrical	Redes Power - Law	0
inst_elisa_126	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_127	Redes Aleatorias	SW	Redes Aleatorias	1
inst_elisa_128	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_129	Redes Aleatorias	SW	Redes Aleatorias	1
inst_elisa_130	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_131	Redes Aleatorias	Geometrical	Redes Power - Law	0
inst_elisa_132	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_133	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_134	Redes Aleatorias	SW	Redes Aleatorias	1
inst_elisa_135	Redes Aleatorias	Geometrical	Redes Power - Law	0
inst_elisa_136	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_137	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_138	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_139	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_140	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_141	Redes Aleatorias	SW	Redes Aleatorias	1
inst_elisa_142	Redes Aleatorias	Geometrical	Redes Power - Law	0
inst_elisa_143	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_144	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_145	Redes Aleatorias	Kleinberg	Redes Aleatorias	1
inst_elisa_146	Redes Aleatorias	SW	Redes Aleatorias	1
inst_elisa_147	Redes Aleatorias	Kleinberg	Redes Aleatorias	1

Continuación Tabla g1. Resultados de la clasificación de redes complejas de naturaleza desconocida.

inst_elisa_148	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_149	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_150	Redes Power - Law	Geometrical	Redes Power - Law	1
inst_elisa_151	Redes Aleatorias	SW	Redes Aleatorias	1
inst_elisa_152	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_153	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_154	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_155	Redes Aleatorias	Caveman	Redes Aleatorias	1
inst_elisa_156	Redes Power - Law	BA	Redes Power - Law	1
inst_elisa_157	Redes Aleatorias	Geometrical	Redes Power - Law	0
inst_elisa_158	Redes Power – Law	BA	Redes Power - Law	1
inst_elisa_159	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
inst_elisa_160	Redes Aleatorias	Aleatorios	Redes Aleatorias	1
			Total de clasificaciones correctas	**146**

ANEXO H

Este anexo contiene los resultados de la comparación realizada entre algoritmos ampliamente utilizados para la selección de características y el enfoque estadístico utilizado para la selección de funciones de caracterización relevantes y no redundantes así como la identificación del conjunto mínimo. Como puede observarse en la sección 3.3, la selección de características tiene dos aspectos importantes a considerar: la búsqueda de subconjuntos en el espacio de características y la evaluación de los atributos de esos subconjuntos. El enfoque estadístico empleado en esta investigación para la selección de características considera estos dos aspectos, las tres primeras etapas de esta actividad (ver Figura 6.6) corresponden a la búsqueda de subconjuntos en el espacio de características y la cuarta etapa corresponde con la evaluación de los atributos del subconjunto.

Dentro del área de Análisis Multivariado podemos encontrar tres métodos para la selección de características, el software utilizado para comparar estos métodos con el enfoque estadístico mencionado anteriormente fue SAS (). En el área de Aprendizaje Automático existe una gran variedad de algoritmos que se incluyen en Weka, una descripción amplia de los evaluadores de atributos y algoritmos de búsqueda utilizados por este programa se puede encontrar en [Witten, 2005]. En la Tabla h1 se muestran los nombres de los métodos y algoritmos para la selección de características empleados en la comparación de los resultados obtenidos con el enfoque estadístico, así como la codificación que será utilizada en la Tabla h2.

Tabla h1. Métodos y algoritmos para la selección de características usados para comparar los resultados obtenidos con el enfoque estadístico.

Codificación	Evaluador de Atributos	Algoritmo de búsqueda
1	*Forward Selection*	
2	*Backward Selection*	
3	*Bidirectional Search*	
4	*CfsSubsetEval*	*BestFirst*

Continuación Tabla h1. Métodos y algoritmos para la selección de características usados para comparar los resultados obtenidos con el enfoque estadístico.

5	*ChiSquaredAttributeEval*	*Ranker*
6	*ClassifierSubsetEval*	*Genetrc*
7	*ClassifierSubsetEval*	*RandomSearch*
8	*ClassifierSubsetEval*	*RankSearch*
9	*ConsistencySubsetEval*	*BestFirst*
10	*ConsistencySubsetEval*	*RandomSearch*
11	*GainRatioAtributeEval*	*Ranker*
12	*InfoGainAtributeEval*	*Ranker*
13	*OneRAtributeEval*	*Ranker*
14	*ReliefFAtributeEval*	*Ranker*
15	*SMVAtributeEval*	*Ranker*
16	*SymetricalUncertAttributeEval*	*Ranker*
17	*SymetricalUncertAttributeSetEval*	*FCBFSearch*
18	*WrapperSubsetEval*	*GeneticSearch*
19	*WrapperSubsetEval*	*RandomSearch*
20	*WrapperSubsetEval*	*RankSearch*

En la Tabla h2 se muestran las funciones de caracterización que fueron seleccionadas con los métodos y algoritmos de la Tabla h1. En la primera columna se enlistan las funciones de caracterización analizadas, en la segunda columna se indican qué funciones de caracterización fueron seleccionadas como características relevantes y no redundantes mediante el enfoque estadístico descrito en este trabajo de investigación; después de estas dos columnas siguen veinte columnas que representan a cada uno de los métodos y algoritmos de la tabla h1, en cada una de ellas se indican qué funciones de caracterización fueron seleccionadas, en el caso de los algoritmos que utilizan tablas de posiciones (Ranker) se indican sólo aquellas características que ocupan las cuatro primeras posiciones. La penúltima columna indica el número de ocurrencias obtenidas de la suma de los algoritmos que seleccionan esa función, en base a este indicador en la última columna

aparecen las funciones de caracterización en orden ascendente en función del número de ocurrencias.

Se puede observar en la última columna de la Tabla h2, que las funciones de caracterización que ocupan las primeras cuatro posiciones coinciden con las funciones de caracterización seleccionadas con el enfoque estadístico, así mismo se observa que la función de caracterización identificada como el conjunto mínimo, ocupa la primera posición en los algoritmos que utilizan Ranker como algoritmo de búsqueda y es seleccionada por los demás algoritmos y métodos.

Tabla h2. Comparación de los resultados de la selección de características efectuada mediante métodos de análisis multivariado, algoritmos de aprendizaje automático y el enfoque estadístico propuesto.orden

Funciones de caracterización	Selección de características																					Número de ocurrencias	Orden según ocurrencias
	Enfoque estadístico	Análisis Multivariado			Aprendizaje Automático																		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
$\langle k \rangle$		✓	✓	✓	✓							2					2					6	6
$\sigma_{\langle k \rangle}$	✓	✓	✓	✓	✓			✓			✓					2				✓		8	**4**
$DDC(G)$	✓	✓	✓	✓	✓	1	✓	✓	✓	✓	✓	1	1	1	1	1	1	✓	✓	✓	✓	20	**1**
$L(G)$	✓	✓	✓	✓		4		✓			✓		2	2	4		4			✓		11	**2**
$D(G)$												3			2	3						3	8
E_{loc}	✓	✓	✓	✓	✓	2		✓			✓			4		4				✓		10	**3**
E_{glob}		✓	✓	✓								4	3	3	3		3					8	5
$CG(G)$		✓	✓	✓		3							4									5	7

REFERENCIAS

[Airoldi 2005] Airoldi, E.M., Carley, K.M.: Sampling Algorithms for Pure Network Topologies: A Study on the Stability and the Separability of Metric Embeddings. ACM SIGKDD Explorations Newsletter. Vol. 7, No. 2. (2005). 13 – 22.

[Albert 2000] Albert, R., Jeong, H., Barabási, A.L.: Error and Attack Tolerance of Complex Networks. Nature. Vol. 506. (2000). 378-382

[Albert 2002] Albert, R., Barabási, A.L.: Statistical Mechanics of Complex Networks. Reviews of Modern Physics. Vol. 74, No.1. (2002). 47 – 97.

[Ali 2004] Ali, W., Mondragon, R.J., Alavi, F.: Extraction of Topological Features from Communication Network Topological Patterns Using Self-organizing Feature Maps. arXiv:cs/0404042v2. (2004).

[Amaral 2000] Amaral, L.A.N., Scala, A., Barthelemy, M., Stanley, H.E.: Classes of Small World Networks. Proceedings of the National Academy of Sciences. Vol. 97, No. 21. (2000) 11149 – 11152.

[Amaral 2001] Amaral, L.A.N., Gopikrishnan, P., Matia, K., Plerou, V., Stanley, H.E.: Application of Statistical Physics Methods and Concepts to the Study of Science and Tecnology Systems. Scientometrics. Vol. 52, No. 1. Budapest. (2001)

[Amaral 2004] Amaral, L.A.N., Ottino, J.M.: Complex Systems and Networks: Challenges and Opportunities for Chemical and Biological Engineers. Chemical Engineering Scientist Vol. 59 (2004) 1653 – 1666.

[Barabási 1999a] Barabási, A.L., Albert, R.: Emergence of Scaling in Random Networks. Science. (1999). 509 – 512.

[Barabási 1999b] Barabási, A.L., Albert R., Jeong, H.: Mean-Field theory for Scale-free Random Networks. Physica A. Vol. 272. (1999). 173 – 189.

[Barabási 2003] Barabási, A.L.: Emergence of Scaling in Complex Networks. Handbook of Graphs and Networks. Wiley-VCH. Berlin. (2003). 69 – 82.

[Barabási 2005] Barabási, A.L.: Taming Complexity. Nature Physics. Vol. 1. (2005).68 – 70.

[Bollobás 2002] Bollobás, B., Riordan, O.M.: Mathematical Results on Scale-free Random Graphs. Handbook of Graphs and Networks. Wiley-VCH. Berlin. (2002). 1 – 32.

[CAIDA 2005] Cooperative Association for Internet Data Analysis.: http://www.caida.org/home/

[CASOS 2005] Center for Computational Analysis of Social and Organizational Systems.: http://www.casos.cs.cmu.edu/

[Costa 2007] Costa, L., Rodrigues, F.A., Travieso, G., Villas, P.R.: Characterization of Complex Networks: A Survey of Measurements. arXiv:cond-mat/05050485v5. (2007).

[Duda 2000] Duda, R.O., Hart, P.E., Strok, D.G.: Pattern Classification. Second Edition. Wiley Interscience. (2000).

[Fabrikant 2002] Fabrikant, A., Koutsoupias, E., Papadimitriou, Ch. P.: Heuristically Optimized Trade-offs: A New Paradigm for Power Laws in the Internet. STOC 02 (2002).

[Faloutsos 1999] Faloutsos, M., Faloutsos, P., Faloutsos, C.: On Power-Law Relationship on the Internet Topology. ACM SIGCOMM Computer Communication Review. Vol. 29, Issue 4. (1999). 251 – 262.

[Goh 2002] Goh, K., Oh, E., Jeong, H., Kahng, B., Kim, D.: Classification of Scale-free Networks. Proceedings of the National Academy of Sciences. Vol. 99, No. 20. (2002) 12583 – 12588.

[Gustafsson 2006] Gustafsson, M., Hörnquist, M., Lombardi, A.: Comparison and Validation of Community Structures in Complex Networks. Physica A. Vol. 367. (2006). 559 – 576.

[Guyon, 2003] Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research. Vol. 3. (2003) 1157 – 1182.

[Hayes 2000] Hayes, B.: Graph Theory in Practice: Part I. American Scientist Vol. 88 No. 1 (2000) 9 – 13.

[Hayes 2000b] Hayes, B.: Graph Theory in Practice: Part II. American Scientist Vol. 88 No. 2 (2000) 104 – 109.

[Hinkelmann 2005] Hilkelmann K., Kempthrone O.: Design and Analysis of Experiments. Vol. 2. Advanced Experimental Design. Wiley Series in Probability and Statistics. (2005).

[Jonson 2000] Jonhson, D.E.: Métodos Multivariados Aplicados al Análisis de Datos. International Thomson Editores. (2000).

[Kecman 2001] Kecman V.: Learning and Soft Computing. Support Vector Machines, Neural Networks and Fuzzy Logic Models. The MIT Press. (2001).

[Latora 2001] Latora, V., Marchiori, M.: Efficient Behavior of Small World Networks. Physical Review Letters. Vol. 87, No.19. (2001).

[Lenth 2006] Lenth R.V.: Java Applets for Power and Sample Size. Computer Software. http://www.stat.uiowa.edu/~rlenth/Power

[Liu 2003] Liu, Z., Lai, Y., Ye, N., Dasgupta, P.: Connectivity Distribution and Attack Tolerance of General Networks with Both Preferential and Random Attachments. Physics Letters A. Vol. 303, Issue. (2003). 337 – 344.

[Manson 2033] Manson, R.L., Gunst R.F., James L.H.: Statistical Design and Analysis of Experiments with Applications to Engineering and Science. Second Edtion. Wiley – Interscience. (2003).

[Michie 1994] Michie, D., Spiegelhalter, D.J., Taylor, C.C.: Machine Learning, Neural and Statistical Classificartion. Introduction. (1994). 1 – 5.

[Middendorf 2004] Middendorf, M., Ziv, E., Adams, C., Hom, J., Koytcheff, R., Levovitz, C., Woods G., Chen, L., Wiggins, C.: Discriminative Topological Features Reveal Biological Network Mechanisms. BMC Bioinformatics 2004. Vol 5, No. 181 (2004).

[Mitchell 1997] Mitchell, T.M.: Machine Learning. McGraw Hill. (1997)

[Montgomery, 2004] Montgomery, D.C.: Diseño y Análisis de Experimentos. Limusa Wiley. (2004).

[Newman 2002] Newman, M.E.J.: Random Graphs as Models of Networks. Handbook of Graphs and Networks. Wiley-VCH. Berlin. (2002). 35-65.

[Newman 2003] Newman, M.E.J.: The Structure and Function of Complex Networks. SIAM Review. Vol. 45. No. 2. (2003). 167 – 256.

[Newman 2006] Newman M.E.J., Barabási A.L., Watts, D.J.: The Structure and Dynamics of Networks. Princeton University Press. (2006).

[Novaes 2005] Novaes De Santana, C.: Analise da Pluviometria do Nordeste Brasileiro Segundo Modelagem em Redes. Monografía. (2005).

[NWB 2005] Network Workbench.: https://nwb.slis.indiana.edu/community/?n= Main.HomePage

[Ortega 2005] Ortega, R.: Estudio de las propiedades topológicas en redes complejas con diferente distribución del grado y su aplicación en la búsqueda de recursos

distribuidos. Propuesta de Tesis de Doctorado. Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada (CICATA). Altamira, Tamaulipas, México. (2005)

[Russell 1996] Russell, S., Norving, P.: Inteligencia Artificial. Un Enfoque Moderno. Prentice Hall. México. (1996).

[Salazar 2005] Salazar, A.M.A.: Pronostico de Demanda por Medio de Redes Neuronales Artificiales (RNA's) en la Industria de Telecomunicaciones. Tesis de Maestría. Universidad Autonoma de Nuevo León. Facultad de Ingeniería Mécanica y Eléctrica. San Nicolas de los Garza, Nuevo León, México. (2005)

[Schmidt 1991] Schmidt, S.R., Launsby, R.S.: Understanding Industrial Designed Experiments. 3rd Edition. Air Academy Press. (1991)

[Sen 2003] Sen, P., Dasgupta, S., Chatterjee, A., Sreeram, P.A., Mukherjee, G., Manna, S. S.: Small-world Properties of the Indian Railway Network. Physical Review E. Vol 67. (2003).

[Servente 2002] Servente, M.: Algoritmos TDIDT Aplicados a la Minería de Datos Inteligente. Tesis de grado en Ingeniería Informática. Facultad de Ingeniería. Universidad de Buenos Aires. (2002).

[Singhi 2006] Singhi, S.K., Liu, H.: Feature Subset Selection Bias for Classification Learning. Proceedings of the 23rd ICML. ACM International Conference Proceeding Series. Vol. 148. (2006). 849 – 856

[Virtanen 2003] Virtanen, S.: Properties of Nonuniform Random Graphs Models. Research Report 77. Helsinkin University of Technology. Laboratory of Theoretical Computer Science. (2003).

[Witten 2005] Witten, H.I., Frank, E.: Data Mining. Practical Machine Learning Tools and Techniques. Second Edition. Elsevier. (2005).

[Yu 2004] Yu, L., Liu, H.: Efficient Feature Selection via Analysis of Relevance and Redundacy. Journal of Machine Learnig Research 5. (2004) 1205 – 1224.

[Zhao 2007] Zhao H.: Semantic Matching Across Heterogeneous Data Sources. Communications of ACM. Vol. 50 No.1. (2007) 45-50.

[Ziv 2005] Ziv, E., Koytcheff, R., Middendrof, M., Wiggins, C.: Systematic Identification of statistically significant network measures. arXiv:cond-mat/0306610v3. (2005).