CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS
DEL INSTITUTO POLITÉCNICO NACIONAL
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA
SECCIÓN DE COMPUTACIÓN

Predicción de Tráfico en Redes de Telecomunicaciones basado en Técnicas de Inteligencia Analítica

Tesis que presenta

Millan Alonso Gildardo

Para obtener el grado de

Maestro en Ciencias

En la especialidad de

Ingeniería Eléctrica

Opción Computación

Director: **Dra. Xiaoou Li Zhang**
Co-director: **Dr. Luis E. Rocha Mier**

México, D.F., Julio de 2006

Agradecimientos

Gracias a mis padres por su ejemplo y apoyo incondicional, cada meta en mi vida es también de ustedes.

A mis hermanos por confiar y alentarme en cada momento.

A mis sobrinos por demostrarme con su cariño que hay alguien por quien seguir adelante.

A ustedes Jorge y Francisco por demostrarme que pudo contar con su amistad en todo momento.

A mis asesores el Dr. Luís E. Rocha y Dra. Xiaoou Li por aconsejarme y transmitirme sus conocimientos y por haberme permitido desarrollar un trabajo que me dejo valiosas enseñanzas.

A mis compañeros de la sección de computación por hacer mas amena mi estancia durante estos dos años.

A Sofi por contar con tu ayuda en todo momento.

Al CINVESTAV por permitirme cursar todas las materias de la maestría.

Al CONACYT por la beca proporcionada para desarrollar este programa de maestría.

Resumen

Durante el proceso de planeación de la capacidad de redes de telecomunicaciones, es necesario estimar el tráfico esperado que la red debería soportar. La estimación en ocasiones es difícil de obtener debido a la naturaleza distribuida de la red y al gran número de variables que se deben de tomar en cuenta.

En este trabajo de investigación, proponemos una metodología basada en técnicas de inteligencia analítica para detectar y pronosticar las variables de red que contribuyen a la utilización de los enlaces. Además, se propone un modelo de redes neuronales en minería de datos capaz de pronosticar el nivel de utilización de los enlaces de una red de datos. Los datos sobre el estado de la red de datos son obtenidos a partir de un simulador de redes de telecomunicaciones y después son analizados gracias a la adaptación de técnicas analíticas. Los resultados experimentales muestran las ventajas de la metodología propuesta para detectar las variables más importantes y la precisión del modelo de predicción para anticipar la carga de tráfico de un escenario de red de datos.

Abstract

During the network capacity planning process, it is necessary to estimate the traffic that the data network links can support. This estimation is sometimes very hard to obtain due to the distributed nature of the network and the great number of network variables.

In this research work, we propose a framework based on analytics intelligence techniques to detect the data network variables that contribute to the traffic load. In addition, we present a data mining predictive model based on neural networks that is able to forecast accurately the most important variables allowing the prediction of links' utilization. The data of the data network behavior are obtained from a telecommunication network simulator, and after they are analysed thanks to the adaptation of analytic techniques. Our experimental results show the advantages of our framework in detecting the most important variables and the accuracy of our developed predictive model to anticipate the traffic load.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1.

Introducción

Las redes de telecomunicaciones generan grandes cantidades de datos operacionales que incluyen tráfico, estadísticas de utilización y datos de alarmas y fallas en varios niveles de detalle. Esta colección de datos regularmente esconde conocimiento e inteligencia que es crucial para algunas tareas involucradas en la administración de una red de telecomunicaciones [19]. Por ejemplo, analizar los datos de las alarmas generadas en una red de telecomunicaciones para encontrar patrones nos permite predecir fallas en componentes de hardware.

La *Planeación de la Capacidad de la Red* (Network Capacity Planning, NCP) es un proceso que modela y simula un gran número de alternativas de red en una organización, incorporando cambios en diseño, tecnologías, componentes, configuraciones, costos y aplicaciones óptimas para proporcionar resultados en términos de desempeño y confiabilidad [2]. Este proceso incluye el pronóstico del tráfico y la identificación de importantes patrones en la red [19]. La NCP puede ayudar a enfrentar problemas antes de que sucedan. Sin embargo, coleccionar los datos de los dispositivos que existen en la red, analizar las variables más importantes y generar un reporte estadístico sobre estas variables y de eventos críticos no es suficiente para predecir el tráfico que la red puede soportar. Además, encontrar información interesante manualmente, por ejemplo con métodos estadísticos, requiere de mayor tiempo [27].

Las redes de telecomunicaciones que generan grandes cantidades de información son candidatas ideales para la aplicación de técnicas de inteligencia analítica [49]. Dentro de esta gran variedad de técnicas se encuentra la minería de datos, que consiste en la exploración y análisis de largas cantidades de datos para descubrir conocimiento [34]. La minería de datos es ideal en redes de telecomunicaciones para:

- Asistir a sistemas que ayuden a la planeación de la capacidad de la red.
- Obtener conocimiento de la gran cantidad de datos que tiene disponibles una red de telecomunicaciones.

- Descubrir de manera automática conocimiento, ya que los expertos en el dominio de las telecomunicaciones a menudo no están enterados de la existencia de patrones en los datos.

- No depender del conocimiento aprendido por los humanos sobre una red de telecomunicaciones. En un futuro, debido a un incremento de las dimensiones de la red o a la incorporación de nuevas aplicaciones, cambiará el conocimiento encontrado, por lo que se necesita constantemente analizar los datos generados por la red.

- Analizar el tráfico de una red de telecomunicaciones no solo se puede basar en el pronóstico de la utilización de los enlaces, sería deseable conocer las variables que están relacionadas y contribuyen con los cambios de los niveles de utilización.

En esta tesis, se maneja el problema de anticipación de la carga de tráfico en los enlaces de redes de datos con la finalidad de mantener su disponibilidad. Este problema es complejo debido a la distribución de la red de datos y el gran número de variables de red que se deben tomar en cuenta. Debido a que las redes de datos frecuentemente contienen miles de componentes y un problema en alguno de ellos puede propagarse rápidamente a través de la red, es muy importante identificar una falla o incremento de tráfico antes que resulte en una falla total en un componente de hardware [46].

Gran parte del problema radica en seleccionar y pronosticar las variables que realmente tienen un efecto en el comportamiento del enlace de red cuyo tráfico se quiere predecir. Por ejemplo, si un canal se bloquea debido al alto tráfico que transita en un momento determinado, sería bueno saberlo antes para evitar que se sobre utilice este canal. Pero sobre todo, sería útil saber cuáles son las variables que contribuyen a elevar el tráfico en un momento determinado.

Para afrontar este problema, se propone una metodología teórica que comprende varios pasos necesarios para el pronóstico de las variables de red y la predicción del tráfico generado de los enlaces de comunicaciones[1].

Actualmente se han desarrollado varios modelos de minería de datos para solucionar el problema de pronóstico de tráfico en una red de datos [5], [31], [10]. Estos modelos se basan en Redes Neuronales (RN) gracias a su eficiente adaptación y buena capacidad de pronóstico en redes de datos. Una RN se concentra en pronosticar el futuro de una

[1]En esta tesis el término pronóstico se refiere a series de tiempo, mientras que el término predicción se refiere a modelos de minería de datos.

serie de tiempo basándose en sus valores pasados [43]. En este trabajo se utilizaron RN con un enfoque diferente, se analizaron estadísticas generadas en una red de datos para detectar variables que estuvieran relacionadas con el nivel de tráfico en la red, y que pudieran sevir como entradas en un modelo de predicción basado en RN.

Una de las contribuciones de esta tesis es describir una metodología que nos permita descubrir conocimiento a partir de las estadísticas de una red de datos.

Esta metodología se puede aplicar a un escenario de red del mundo real para detectar las variables más importantes que están relacionadas con el tráfico, para después constantemente recolectar sus estadísticas para poder predecir el tráfico de alguno de los enlaces de la red.

1.1. Trabajo relacionado

Durante los últimos años ha iniciado la unión de dos áreas muy importantes: telecomunicaciones y minería de datos. Se han propuesto sistemas que desarrollan un proceso de adquisición de conocimiento por medio del análisis de la gran cantidad de datos que proporciona una red de telecomunicaciones. Estos sistemas ayudan a la administración, detección y análisis del comportamiento de las redes de telecomunicaciones.

Los datos que genera una red de telecomunicaciones incluyen detalles de las llamadas, los cuales describen las llamadas que se realizan en la red de telecomunicaciones, datos de la red, los cuales describen el estado de los componentes de software y hardware de la red, datos de los clientes, los cuales describen a los clientes que se les brinda algún servicio de telecomunicaciones.

Varias aplicaciones de minería de datos han utilizado esta gran cantidad de datos para resolver problemas en las redes de telecomunicaciones tales como: identificar fraudes, mejorar el desempeño de marketing, identificar fallas en componentes de hardware y prónosticar el tráfico de la red.

Esta tesis está enfocada en considerar los datos de una red de telecomunicaciones que detallan el estado de los componentes de software y hardware. Y de acuerdo al análisis que se hace sobre este tipo de datos se clasifican los trabajos de investigación en dos grupos:

1. *Trabajos de investigación* que identifican y predicen fallas de dispositivos de hardware, por medio del análisis de alarmas generadas entre los dispositivos de una

red de telecomunicaciones.

Un sistema que permitió la obtención de conocimiento oculto acerca del comportamiento de redes de telecomunicaciones fue el realizado por K. Ratonen, M. Klemettinen y H. Mannila, llamado "Knowledge discovery from telecommunication network alarm databases"[25], el cual describe un sistema llamado TASA (Telecommunication Alarm Sequence Analyzer) que permite descubrir conocimiento a partir de bases de datos de alarmas generadas en redes de telecomunicaciones. Este sistema incorpora dos partes del proceso de KDD[2]: descubrimiento de patrones (minería de datos) y posprocesamiento. El sistema esta basado en una estructura que permite localizar la ocurrencia de episodios (conjunto de alarmas con determinado orden) dentro de una secuencia de datos, este conocimiento descubierto es expresado en términos de reglas. Debido a que TASA es un proceso interactivo e iterativo, brinda al usuario una interfaz en la cual pueda manipular las reglas descubiertas usando operaciones de orden y selección. Su desventaja es que necesita la intervención de una persona experta en el dominio de las telecomunicaciones para que pueda hacer una interpretación de las reglas encontradas, por lo que es un sistema semiautomático.

Durante la evolución de sistemas inteligentes que brinden soluciones a problemáticas en las redes de telecomunicaciones, se encuentrá las aportaciones realizadas por Gary Weiss, investigador de los laboratorios de AT&T, quien ha aportado trabajos como "Intelligent Telecommunication Technologies"[45], donde describe aplicaciones inteligentes en el área de Telecomunicaciones. Otro trabajo realizado por este investigador, fue el desarrollar un sistema experto para AT&T llamado ANSWER (Automated Network Surveillance with Expert Rules) [46], el cual es responsable de monitorear los switches 4ESS que rutean la mayoría del tráfico en la red de AT&T. El switch genera una alarma y la envía a uno de los dos centros de control técnico de AT&T. En el centro de control técnico, la alarma se inserta en una base de datos relacional y lo reenvía a ANSWER, donde se analiza la alarma usando reglas obtenidas del dominio experto. Si ANSWER determina que ante la alarma se tiene que tomar una acción, envía una alerta, describiendo el problema a un técnico para más procesamiento. La tarea de la minería de datos en este caso de estudio es identificar patrones en los registros de las alarmas de red que se puedan usar para predecir fallas en el equipo de telecomunicaciones. Estos

[2]Descubrimiento de Conocimiento de Base de datos (Knowledge Discovery in Databases, KDD) que consiste en una secuencia iterativa de limpieza, integración, selección, transformación, minería y evaluación de datos. [27].

patrones entonces se pueden incorporar como reglas en el sistema ANSWER. El algoritmo genético utilizado para la minería de datos permite encontrar patrones que no ocurren muy frecuentemente en los datos, aunque solo permite predecir fallas de manera correcta, en tiempos de advertencia cortos.

En múltiples trabajos se ha utilizado un proceso de adquisición de conocimiento, pero el que se destaca por describir a detalle una metodología para descubrir conocimiento en la base de datos de alarmas de una red de telecomunicaciones es la tesis de Mika Klemettinen titulada "A knowledge discovery methodology for telecommunication network alarm databases"[27], en la que dedica un par de capítulos para describir el comportamiento de una red de telecomunicaciones y el formato de las alarmas (definir parámetros de la alarma), así como la forma y el tipo de episodios y asociación de reglas a encontrar. Considera que un sistema como TASA (mencionado anteriormente), es adecuado para descubrir conocimiento ya que permite la comprensibilidad, existencia de eficientes algoritmos para la búsqueda de las reglas y permite manipular características de la aplicación (límites de frecuencia y confianza). La metodología que utilizó en el proceso de KDD consiste en los siguientes pasos: preprocesamiento de datos, transformación y selección de datos, descubrir patrones (minería de datos), presentación y posprocesamiento de los resultados y por último la interpretación y utilización de los resultados. Después de este proceso, Mika Klemettinen concluye que las redes de telecomunicaciones usualmente cambian y evolucionan rápidamente, aún con pequeños cambios en ellas se puede afectar algunas veces de manera sustancial su comportamiento. Esto significa, que reglas viejas que se tienen hasta un cierto momento no necesariamente se mantendrán en un ambiente cambiante, por ello es necesario un sistema que sea iterativo y que permita a los humanos decidir si al correr nuevamente el procedimiento de adquisición de conocimiento, requiere modificar las reglas base del sistema experto de correlación.

Diferentes autores han trabajado en conjunción con empresas de telecomunicaciones para el estudio y aplicación de diferentes técnicas de minería de datos en problemas del dominio de las telecomunicaciones. Un ejemplo de ello es el desarrollado por Sterritt, R., Adamson, K., Shapcott, C.M., Curran, E.P, y NITEC (Nothern Ireland Telecommunications Engineering Centre) llamado "Data mining telecommunications network data for fault management and development testing"[42], en el que han aplicado técnicas de minería de datos para TMN (Telecommunications Management Network) con el propósito de administrar fallas y desarrollar pruebas en sistemas de Telecomunicaciones. Este trabajo inicio

con la simulación de componentes de red en una arquitectura paralela para facilitar las pruebas a gran escala, hasta evolucionar a un proyecto que pudiera manejar componentes de red con velocidades de tiempo real, y brindar al usuario una interfaz gráfica que le permitía mapear alguna topología de red deseada a un ambiente paralelo. En la administración de alarmas se monitorea, filtra y enmascara las alarmas para manejar solo las de mayor prioridad. Por medio de la correlación reducen el número de alarmas, y dejan al operador el criterio para determinar la causa.

La importancia de considerar un análisis de alarmas, es que cada alarma contiene datos que fluyen a través de la red y por lo tanto contribuyen con el tráfico generado en algunos de los enlaces de una red de telecomunicaciones.

2. *Un segundo tipo de trabajos de investigación*, son aquellos que se enfocan en analizar los datos generados en los componentes de hardware para pronosticar el tráfico de una red de telecomunicaciones. Ejemplo de este tipo de datos son las estadísticas como: throughput, retardo y utilización de los enlaces de una red de telecomunicaciones, utilización del CPU de los servidores de aplicaciones, tráfico enviado y recibido por los switches y routers de la red.

   Se han aplicado varias técnicas de minería de datos como RN [17], redes Bayesianas [29], modelos de regresión [21] y estadísticos [32] para la pronóstico del tráfico en una red de telecomunicaciones. El tráfico de una red de datos se puede representar por una serie de tiempo. Una serie de tiempo es generalmente una secuencia de medidas de una o más variables de un sistema dinámico.

   En los últimos años se han aplicado modelos de minería de datos basados en RN, para el pronóstico de series de tiempo de tráfico en una red de datos [5], [31]. Estos modelos pronostican el nivel de tráfico en la red basándose en los valores pasados de la serie de tiempo, de esta manera el número de entradas de la Red Neuronal dependerá del tamaño de la ventana de tiempo con el que se realice el pronóstico, es decir si se realiza el pronóstico del tráfico en un tiempo t+1, basándose en las muestras t, t-1, t-2,..., t-N, entonces la red neuronal tendrá N entradas y una salida. Existen modelos que han optimizado el tamaño de la ventana de tiempo [44], [38]. Sumado ha esta investigación se han propuesto modelos que combinan diferentes arquitecturas de redes neuronales para aumentar la eficiencia del nivel de pronóstico [5].

   También se han utilizado modelos de regresión como ARIMA (AutoRegressive Integrated Moving Average) para el pronóstico a largo plazo del tráfico generado

en una red de datos [26]. En este trabajo proponen una metodología que permite pronosticar una serie de tiempo del tráfico generado en una red de datos, donde basan el pronóstico en las estadísticas de los valores pasados del nivel de tráfico. También demuestran que cambios en el ruteo de los datos y en la topología de la red genera cambios en la cantidad de tráfico que se genera en algunos enlaces de la red de datos.

Los trabajos realizados para analizar las variables de una red de datos y poder conocer el comportamiento del tráfico se han enfocado en realizar el pronóstico de dichas variables, basándose en los valores pasados para tratar de pronosticar el futuro. Esta tarea se ha resuelto utilizando RN. Sin embargo sería deseable no solo en conocer el comportamiento futuro de dichas variables, si no el de conocer la relación o dependencia que existe entre ellas.

De esta manera, nuestro trabajo de investigación se concentra en seleccionar las variables más importantes que están relacionadas con los niveles de tráfico en una red de datos y hacer el pronóstico de estas variables para que puedan ser entrada de un modelo de predicción. Este modelo de predicción consiste en tener n variables de entrada que permitan predecir la variable tráfico.

Debido a que el propósito es pronosticar un conjunto de variables de red pero además de conocer la relación o dependencia que existe entre ellas se propone utilizar un modelo Vector Autoregresión (VAR). Este modelo en un inicio se enfoco en detectar relaciones en variables económicas y pronosticar el valor de cada variable basándose en sus propios valores pasados y los valores pasados del resto de las variables [11]. En este trabajo de investigación se propone utilizar el modelo VAR como técnica de pronóstico de variables de una red de telecomunicaciones.

1.2. Organización de la tesis

El contenido de la tesis se organiza de la siguiente manera:

En el capítulo 2 se describen los conceptos básicos sobre redes de telecomunicaciones, los problemas que involucra la administración de una red de datos, así como las variables que definen el comportamiento del tráfico en la red. Se mencionan las herramientas que permiten simular escenarios de redes de datos.

En el capítulo 3 se da una introducción de las técnicas de *inteligencia analitica* y se analizan aquéllas que se han utilizado para el pronóstico de series de tiempo y la predicción de tráfico en redes de telecomunicaciones. Además se mencionan tipos de herramientas para la investigación y desarrollo de técnicas de inteligencia analítica.

En el capítulo 4 se explica la metodología propuesta para la predicción de tráfico. En cada uno de los pasos se mencionan los algoritmos y herramientas utilizadas.

En el capítulo 5 se menciona el caso de estudio aplicando cada uno de los pasos de la metodología propuesta.

En el capítulo 6 se dan las conclusiones de este trabajo, señalando las ventajas y desventajas del modelo de predicción utilizado. Además se dan algunas líneas de trabajo futuro que se podrían seguir a partir de esta tesis.

Capítulo 2.

Redes de Telecomunicaciones

Las redes de telecomunicaciones están creciendo rápidamente en tamaño y complejidad, por tal motivo su administración está llegando a ser muy difícil. Los elementos de la red generan gran cantidad de tráfico, y en gran medida depende del tipo de topología y aplicaciones que se utiliza. La NCP de la red analiza esta gran cantidad de tráfico para poder responder y atender problemas antes de que éstos se presenten. En este capítulo se mencionan las diferentes topologías de una red de datos, el tipo de variables que se tienen que considerar para la predicción del tráfico de la red y la manera en que la planeación de la capacidad de la red ha resuelto problemas en las redes de datos. Además, se menciona software que permite generar tráfico de una red de datos.

2.1. Tipos de redes de telecomunicaciones

Hay dos principales tipos de estructuras de red, redes móviles y redes fijas [27]. Pero sin importar el tipo de red de telecomunicaciones, estas se pueden ver como un conjunto de componentes interconectados: switches, servidores, equipo de transmisión, etc. Cada componente puede tener subcomponentes. El número de componentes utilizados depende del nivel de abstracción usado en la red.

Redes móviles

En la figura 2.1 se muestra un ejemplo de la estructura de una red móvil. Un sistema que administra la red (Network Management System, NMS) controla varios controladores de estaciones móviles (Mobile Station Controller, MSC), el cual consiste de varios controladores de estaciones base (Base Station Controller, BSC) y transreceptores de estaciones base (Base Station Transceiver, BTS). Cada componente mencionado contiene varios subcomponentes y dispositivos físicos. El número de componentes depende del nivel de abstracción usado en el sistema, por ejemplo una red operada por una compañía telefónica de área metropolitana puede consistir de miles de componentes en

diferentes niveles [25].

Cada subcomponente y modulo de software en una red de telecomunicaciones puede producir alarmas. Una alarma es un mensaje emitido por un elemento de red, generalmente cuando ocurre un problema. Las alarmas son mensajes describiendo alguna situación anormal; no necesariamente indican que ha ocurrido un problema en la red que es visible para los usuarios [25].

Desafortunadamente, cada elemento de red tiene solo tiene una visión estrecha de todo lo que ocurre en la red, y solo puede reportar los síntomas de las fallas de la visión que tiene. Además, una falla puede resultar en un número de diferentes alarmas de varios elementos de red. En la figura 2.1 adaptada de [25], se muestra que las alarmas se rutean de elementos de bajo nivel hasta el sistema de administración de la red.

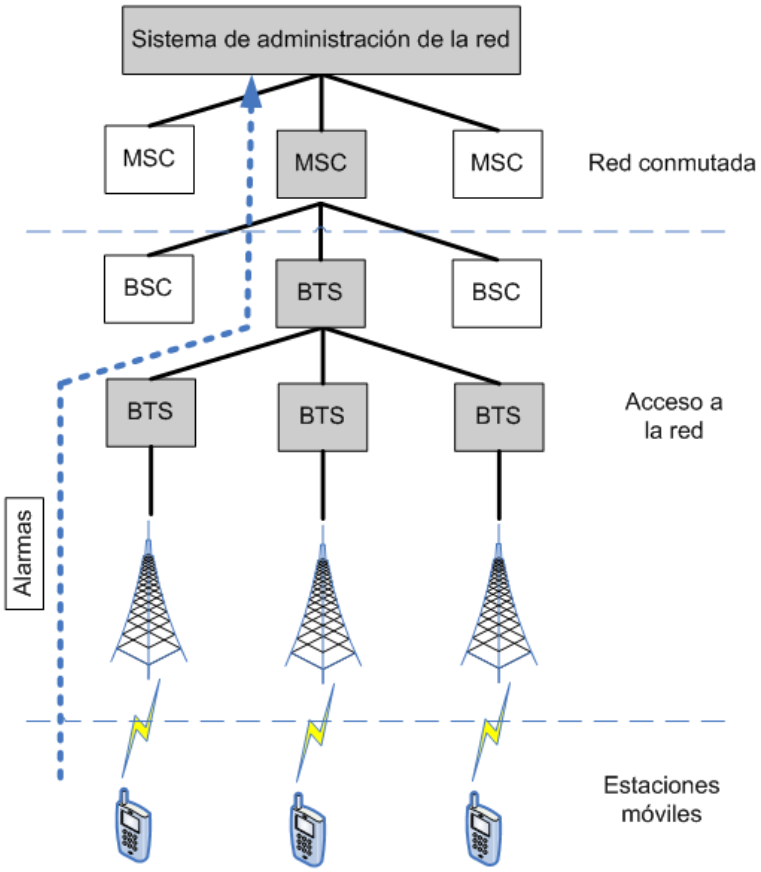

Figura 2.1.: Estructura de una red móvil.

Redes fijas

Si solo se quiere la comunicación entre dos computadoras que se encuentran en la misma oficina, entonces la transmisión consiste solo en un enlace punto a punto, como se muestra en la figura 2.2 (a). Sin embargo, si están localizadas en diferentes partes de un pueblo o ciudad, se deben utilizar servicios públicos. Normalmente se refiere a una red PSTN (Public Switched Telephone Network) la cual requiere un dispositivo conocido como modem para la transmisión de los datos, como se muestra en la figura 2.2 (b).

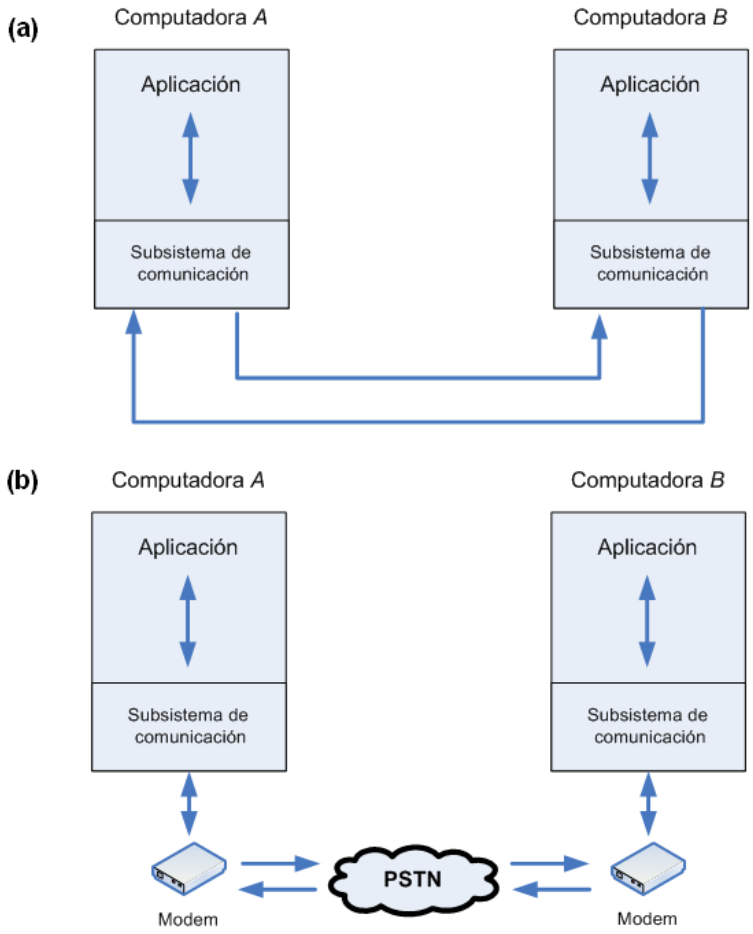

Figura 2.2.: Comunicación entre dos computadoras. (a) Enlace punto a punto (b) PSTN + enlace.

Cuando más de dos computadoras intervienen en una aplicación se utiliza una red de

conmutación que permita comunicar a todas las computadoras en diferentes tiempos. Si todas las computadoras están distribuidas en una sola oficina o edificio, es posible instalar una red propia. Tal red se conoce como redes de área local (Local Area Network, LAN), ejemplo de este tipo de red se presenta en la figura 2.3.

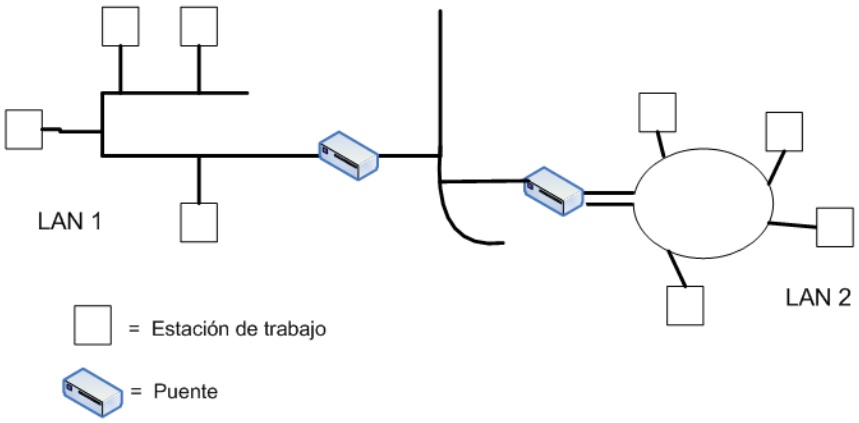

Figura 2.3.: Red de área local (LAN).

Cuando las computadoras están localizadas en diferentes establecimientos (sitios), se deben ocupar servicios públicos. Este tipo de red se conoce como redes de área amplia (Wide Area Network, WAN). El tipo de redes WAN depende de la naturaleza de la aplicación. Por ejemplo, si todos los componentes pertenecen a una misma empresa es simplemente contratar líneas de transmisión (circuitos) de algún proveedor e instalar un sistema de conmutación privado en cada sitio, este tipo de red se llama red privada (Enterprise Private Network). Muchas empresas eligen hacer esto, donde generalmente en estas redes incorporan comunicaciones de voz y datos, un ejemplo de esta red se muestra en la figura 2.4.

Este tipo de soluciones solo son viables para grandes empresas debido a que se debe generar el suficiente tráfico para justificar el costo de la renta de los enlaces, la instalación y administración de una red privada. Varios proveedores de servicios (public carriers) ahora proporcionan un servicio de datos. Estas redes, al igual que las PSTN, son conectadas internacionalmente y han sido diseñadas específicamente para la transmisión de datos. Por tal motivo, para aplicaciones que involucran computadoras distribuidas alrededor de una ciudad o incluso de otro país, se usa una red PSDN (Public Switched Data Network). Muchos proveedores de servicios han convertido sus redes PSTN para que puedan transmitir datos sin utilizar modems. Como resultado de este

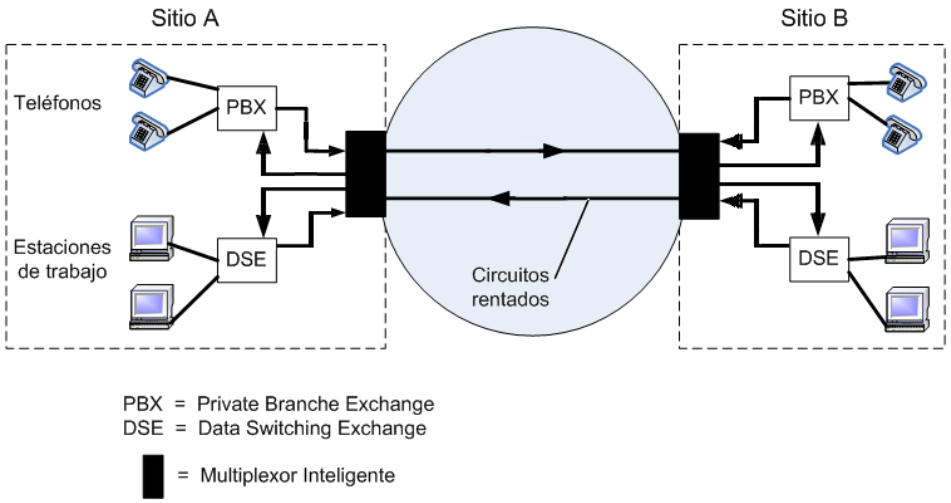

Figura 2.4.: Red de área amplia (WAN).

cambio están las redes digitales de integración de servicios (Integrated Services Digital Networks, ISDN), las cuales operan en un modo totalmente digital.

Hasta ahora se ha considerado que las computadoras pertenecen a la misma LAN o WAN. En muchas aplicaciones sin embargo, la comunicación de los datos involucra la interconexión de múltiples redes como LAN-WAN-LAN. Por ejemplo, una estación de trabajo (computadora) que pertenece a una red LAN en un edificio de una empresa se puede comunicar con otra computadora que pertenece a otra LAN, por medio de la conexión de ambas redes LAN a una red PSDN, como se muestra en la figura 2.5. Este tipo de conexión se conoce como Internet [23].

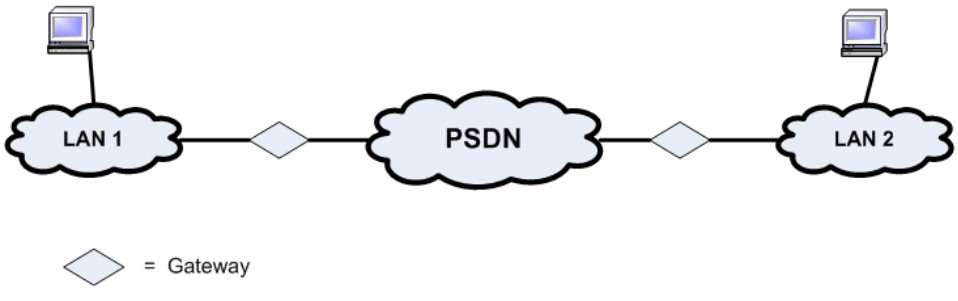

Figura 2.5.: Internet.

Las redes que se han mencionado se han diseñado principalmente para la transmisión de datos entre dos estaciones de trabajo que soportan servicios solo de datos. Pero en

los últimos años, las estaciones de trabajo se han utilizado para soportar servicios que involucren además de datos otros tipos de información como videoconferencia, telefonía con video y servicios multimedia.

2.2. Tráfico en una red de telecomunicaciones

El tráfico de red consiste en la cantidad de datos que fluye a través de la red. Los datos que genera una red de telecomunicaciones incluyen detalles de las llamadas, los cuales describen las llamadas que se realizan en la red de telecomunicaciones, datos de la red, los cuales describen el estado de los componentes de software y hardware de la red, datos de los clientes, los cuales describen a los clientes que se les brinda algún servicio de telecomunicaciones [48].

Las redes de telecomunicaciones son configuraciones complejas de miles de componentes interconectados. Cada elemento de la red es capaz de generar datos sobre su comportamiento y el tipo de tráfico que fluye a través de él. Como se menciono en la Introducción, este tipo de datos pueden ser:

1. Alarmas generadas por dispositivos de hardware. Una alarma esta definida por un conjunto de variables que describen el comportamiento del problema que representan, ejemplo, tiempo (instante en que se generó la alarma), identificador (identificador del elemento de hardware o software que origino la alarma) y mensaje (información acerca del problema). La alarma contiene toda la información acerca del problema. La información contenida en las alarmas varía mucho. Algunas alarmas reportan problemas en conceptos lógicos, como caminos virtuales, algunas sobre dispositivos físicos, como perdida de potencia. Aún en redes de telecomunicaciones pequeñas hay muchas alarmas, se generan alrededor de 200-10 000 alarmas por día [27].

2. Estadísticas generadas por los componentes de hardware de la red. Una red de telecomunicaciones se puede ver como un conjunto de componentes interconectados: switches, routers, equipo de transmisión, servidores, etc. Cada dispositivo o modulo de software en una red de telecomunicaciones puede producir estadísticas acerca del comportamiento del tráfico de la red.

Esta gran cantidad de datos se pueden analizar para resolver algunas de las tareas que involucra la administración de una red de telecomunicaciones. Si la finalidad es conocer la forma en que estos datos contribuyen con la cantidad de tráfico que fluye por los dispositivos de la red, es necesario mencionar cuales son las causas y variables

que están relacionadas con la generación del tráfico.

Tipo de tráfico

La generación del tráfico se debe principalmente a tres fuentes de datos: los dispositivos de la red de telecomunicaciones, el número de usuarios a los que se les brinda un servicio y la variedad de servicios que proporciona la red.

Los componentes de una red de telecomunicaciones generan constantemente datos entre ellos para mantener su configuración, sincronización y comunicación. Dentro de este tipo de datos se encuentran la gran cantidad de alarmas generadas para reportar un error en algún dispositivo.

Cada usuario dentro de una red de telecomunicaciones genera una cantidad de tráfico de acuerdo a los servicios que solicita. Algunos de los servicios que pueden generar el tráfico en una red de datos son:

- DHCP. Tráfico generado durante adquisición, renovación y liberación de direcciones IP;
- Validación. Tráfico generado por la autenticación de un usuario en la red;
- Sesiones. Tráfico generado cuando dos computadoras inician una sesión;
- Navegación por Internet. Tráfico generado por las aplicaciones de un navegador de Internet para descargar páginas de un sitio de red.

Debido a que cada componente genera una determinada cantidad de datos, el nivel de tráfico aumentará conforme lo hagan las dimensiones de la red. En un inicio las redes fueron planteadas para dar servicios de transferencia de datos, con el tiempo fueron requeridos nuevos servicios como videoconferencia, aplicaciones web y voz sobre IP que hacen que el nivel de tráfico aumente considerablemente.

Cada día son más las empresas y usuarios que requieren los servicios de una red de telecomunicaciones para hacer frente a una gran variedad de problemas, por tal motivo aumentará el número de solicitudes de servicios a la red, provocando un incremento en los niveles de tráfico.

Variables de tráfico

Como mencionamos anteriormente el tráfico generado en la red dependerá del tipo de aplicaciones y servicios que se proporcione, de tal manera que sería deseable conocer las variables relacionadas con cada uno de los servicios. Sin embargo hay variables que se tienen que considerar sin importar el tipo de aplicación que se tenga en la red.

Una de las variables a considerar para conocer el nivel de tráfico que fluye a través de la red es midiendo el *nivel de utilización* de los enlaces entre las redes de datos, esta variable nos puede indicar el porcentaje de utilización del enlace de transmisión. La utilización de un enlace esta definido por la siguiente ecuación:

$$(\,\%)utilizacion = 100 * \frac{th}{v} \tag{2.1}$$

donde:
v es la *velocidad del enlace* (bits/sec),
th es el *throughput del enlace* (bits/sec).

Para conocer el *nivel de utilización* es necesario calcular el *throughput* que consiste en la tasa de transferencia de la línea de comunicación. Esta variable es una buena medida de la capacidad del canal de un enlace de comunicación y nos permite estimar realmente cuantos bits pasan por segundo. El *throughput* se calcula con la siguiente ecuación:

$$th = \frac{l}{\frac{d}{v} + \frac{l}{r}} \tag{2.2}$$

donde:
d es la *distancia entre los componentes que se comunican*,
v es la *velocidad de propagación de la señal*,
l es la *longitud de la trama* (bits).

Para conocer realmente la cantidad de bits por segundo que se transmiten, el throughput considera el retardo en una transmisión. El retardo puede ser ocasionado por la distancia entre los componentes que se comunican, la codificación, decodificación, señalización, etc. La ecuación 2.2 solo considera el retardo de la transmisión ocasionado por la distancia entre los componentes.

Además de las variables involucradas en los enlaces de comunicación es importante considerar aquellas que nos indican el desempeño de cada componente de hardware.

Algunas de las variables que nos indican el desempeño de un dispositivo son las siguientes:

- Porcentaje de utilización del CPU.
- Tráfico enviado y recibido (bits/sec).
- Tiempo de procesamiento de las tareas (sec).
- Carga de trabajo (solicitudes/sec) y (sesiones/sec).

Si un canal es sobrecargado se puede perder la comunicación entre dos componentes de la red, por tal motivo es importante considerar el nivel de utilización de los enlaces cuando se lleva acabo un análisis del tráfico generado en la red. Es por eso que en este trabajo de investigación el objetivo principal es predecir el nivel de utilización de un enlace en una red de datos.

Un proceso que puede analizar el comportamiento de una red de telecomunicaciones para ayudar a su administración es la NCP.

2.3. Planeación de la capacidad de la red

Uno de los principales cambios a los que se tiene que afrontar un diseñador de red es garantizar soluciones de diseño que proporcionen los resultados esperados en términos de desempeño y confiabilidad. Frecuentemente, una suposición incorrecta en un diseño puede ocasionar resultados catastróficos en la implementación. La pregunta que se hacen los administradores de red es si la red que administran fue diseñada para hacer frente a las nuevas dimensiones y servicios de la empresa.

La NCP es un proceso que modela y simula un gran número de alternativas de red en una organización, incorporando cambios en diseño, tecnologías, componentes, configuraciones, costos y aplicaciones optimas para proporcionar resultados en términos de desempeño y confiabilidad [2].

La NCP es un proceso que los administradores de red utilizan para planear el crecimiento de la red y asegurar que los dispositivos de red que se utilizan puedan responder a los requerimientos de la expansión. La NCP involucra muchos sistemas de red, incluyendo servidores, componentes de red como switches, enlaces, routers, etc. También involucra planear la capacidad de la red durante diferentes periodos de tiempo por

medio de la anticipación del tráfico y la identificación de patrones en el tráfico de la red [36].

La NCP involucra un análisis de las velocidades y capacidades de los enlaces de la red. Esto se puede realizar por medio del análisis de los reportes y datos de estadísticas que se producen el los servidores o en algún otro dispositivo de la red. Sin la NCP los administradores deben de afrontar los problemas de red conforme éstos ocurran, generalmente respondiendo a alguna alarma que indica que hay una posible falla en la red. La NCP puede ayudar afrontar problemas antes de que ocurran.

La importancia de la NCP es mantener la disponibilidad de la red de datos. La red de datos es el medio de comunicación de una empresa, por lo tanto si deja de funcionar la empresa también lo hará, por ello es mejor anticipar los problemas que simplemente resolverlos cuando se presenten.

Existen varias empresas dedicadas a la NCP así como una gran variedad de herramientas para analizar el tráfico. Un ejemplo de una metodología utilizada para llevar acabo la NCP incluye los siguientes pasos [2]:

1. *Definir métricas y requerimientos del proyecto.* Consiste en definir el tipo de variables que se tienen que analizar, por ejemplo: ancho de banda, protocolos de transporte que maneja la red, características de la red, características de administración, calidad de servicio. Se establecen criterios para evaluar las diferentes alternativas de diseño.

2. *Analizar las topologías de la red.* Validar la información en la topología de la red por medio del análisis de mapas, diagramas y documentación de la red.

3. *Capturar tráfico de la red.* Después de entender por completo la topología de la red, se tiene que coleccionar el tráfico de la red de algunos dispositivos relevantes dentro de una red LAN o WAN.

4. *Crear y validar el modelo de red.* El tráfico y la topología de red obtenidos, se pueden utilizar para generar un modelo que se pueda simular en alguna herramienta de simulación de redes de telecomunicaciones, para conocer la forma en que la red se comportaría con diferentes cargas de tráfico.

5. *Evaluar alternativas de diseño.* Ayudar analizar los resultados de la simulación y elegir las más apropiadas alternativas de diseño.

2.4. Discusión

En este capítulo se abordaron algunos de los conceptos necesarios para entender el comportamiento de una red de telecomunicaciones. Los datos que se analizan en este trabajo de investigación se obtuvieron de una red WAN, debido a que permite la interconexión de varias redes LAN con el propósito de brindar servicios a un determinado número de usuarios. Como se menciono anteriormente debido a la importancia de entender y predecir el tráfico para evitar fallas en componentes de hardware, es necesario analizar la variable de utilización de los enlaces en una red de datos. También se hablo sobre los problemas a los que se afronta un administrador de red, y que puede hacer frente gracias a la NCP. Uno de los objetivos de este trabajo de investigación es poder ayudar a la NCP por medio de la predicción del tráfico en una red de datos.

Para hacer frente a la gran cantidad de tareas que involucra la administración de una red de telecomunicaciones, y en especial a la predicción de tráfico se pueden utilizar técnicas de inteligencia analítica.

Capítulo 3.

Técnicas de Inteligencia Analítica

La inteligencia analítica analiza los datos existentes de una empresa para dar soporte a la toma de decisiones y a la integración de un ambiente de negocios inteligente [1]. En este capítulo se mencionan los diferentes modelos que integran la inteligencia analítica y en especial a la minería de datos. Se introduce las tareas que se puede aplicar la minería de datos, y se explica la manera en que una red neuronal se puede aplicar para resolver las tareas de pronóstico y predicción de variables. Además, se explican algunas técnicas estadísticas utilizadas en esta tesis para el procesamiento de los datos.

3.1. Inteligencia analítica

La inteligencia analítica ha sabido dar respuesta a la gran cantidad de preguntas que se generan los administradores de una empresa, por ejemplo: ¿como se puede anticipar el futuro para poder actuar de manera preactiva en lugar de reactiva?, ¿Cómo se puede compartir el conocimiento estadístico de múltiples fuentes de datos?, ¿Cómo integrar conocimiento en sistemas operacionales?, etc.

La inteligencia analítica proporciona una amplia gama de procesos para predecir y describir modelos, pronóstico, optimizaciones matemáticas, simulación, diseños experimentales y otras capacidades analíticas [1]. Estas capacidades se pueden aplicar a un gran número de problema de negocios.

La inteligencia analítica permite transformar preguntas en modelos que se puedan probar y validar, proporcionando sólidas respuestas que obtiene de los datos. Esta información que se obtiene le da a una empresa la capacidad de responder mejor y actuar de manera adecuada y efectiva.

Componentes de la inteligencia analítica

La inteligencia analítica proporciona a una empresa las más amplia gama de algoritmos y capacidades matemáticas de modelado y manipulación de datos. Sus componentes se pueden agrupar de la siguiente manera:

- Soluciones de minería de datos. Esta integrado por un conjunto de algoritmos de minería de datos (RN, árboles de decisión, regresión, clustering) que permiten entender el conocimiento de los datos para después integrarlo a un sistema operacional.
- Métodos robustos de optimización y estadística. Incluye análisis de varianza, regresión, datos categóricos, cluster y selección de muestras de medición. Ayuda a los usuarios al análisis de datos y poder hacer informes para la toma de decisiones.
- Métodos de pronóstico y análisis econométrico. Permite el análisis de series de tiempo, es decir entender datos pasados para pronosticar el futuro o mejor aun entender como una empresa puede responder a los problemas antes que sucedan. Detecta factores que afectan a una empresa, como indicadores económicos o condiciones de mercado, que después se pueden integrar al análisis econométrico y de pronóstico.

Para hacer frente al problema de predicción de tráfico en una red de datos es necesario considerar varias de las técnicas que involucra la inteligencia analítica. Por ejemplo para obtener información del tráfico es necesario utilizar técnicas estadísticas de correlación y selección. Para desarrollar un modelo de pronóstico y predicción es necesario utilizar algunas de las técnicas aplicadas a las tareas de minería de datos.

3.2. Técnicas estadísticas

Una estadística se refiere a medidas tomadas de una muestra de datos [34]. Muchos de los datos usados en minería de datos son discretos por naturaleza. Datos discretos aparecen en la forma de productos, canales, regiones e información descriptiva de negocios. A continuación se hablará sobre la representación que pueden tener los datos discretos.

3.2.1. Series de tiempo

Una estadística descriptiva sobre datos discretos es el número de veces que un valor puede ocurrir, ejemplo de esto es un histograma. Sin embargo, un histograma describe

un solo momento y la minería de datos esta frecuentemente relacionada con lo que esta pasando con respecto al tiempo [34].

Una serie de tiempo es una secuencia de vectores, $\boldsymbol{x}(t)$, t= 0,1,..., donde t representa un lapso de tiempo. Teóricamente, x puede ser un valor que varia continuamente con respecto a t, un ejemplo es el nivel de utilización de los enlaces en una red de datos. En la práctica, para cualquier sistema físico, x será muestreado dado una serie de puntos de datos discretos, *igualmente espaciados en el tiempo*. Las series de tiempo son generalmente secuencias de medidas de una o más variables de un sistema dinámico [38].

Se dice que una serie de tiempo puede descomponerse en cuatro componentes que no son directamente observables, de las cuales únicamente se pueden obtener estimaciones. Estos cuatro componentes son:

1. *Tendencia* (T). Representa los movimientos de larga duración, también se le conoce como evolución subyacente de una serie.

2. *Ciclo* (C) caracterizado por oscilaciones alrededor de la tendencia.

3. *Estacionalidad* (S). Corresponde a fluctuaciones periódicas de la variable, en periodos relativamente cortos de tiempo.

4. *Irregularidad* (I) son movimientos erráticos que no siguen un patrón específico y que obedecen a causas diversas.

El tipo de estadísticas que se analizan en este trabajo de investigación son aquellas que se generan en dispositivos de hardware de una red de datos y debido a que son valores discretos de una variable es posible hacer una representación de serie de tiempo con ellas.

3.2.2. Correlación de Spearman

Para la minería de datos es importante conocer los datos. Y debido principalmente a que algunos sistemas son demasiadas las variables que se tienen que considerar, es importante identificar aquellas variables que no sean redundantes para disminuir el número de variables que se tienen que analizar.

La correlación es una técnica que se puede utilizar para resolver este problema. En este trabajo de investigación se utilizó la correlación de Spearman, esta técnica mide el grado de relación entre dos variables cuantitativas. Las ventajas que presenta son:

- Hacer la correlación de variables intervalo.
- Aplicable para variables que no tienen una distribución normal de sus datos.
- El tamaño de las muestras de datos puede ser pequeño.

El principal motivo del porque utilizar la correlación de Spearman y no alguna otra como la de Pearson, es que no se conoce la distribución de los valores de las variables que componen la serie de tiempo. Si se garantizara una distribución normal en las variables, se podría aplicar la correlación de Pearson.

La correlación de Spearman se calcula de la siguiente manera:

1. Ordenar los datos de ambas variables. Al valor más pequeño asignar el valor de 1.
2. Asignar un número a cada dato, de acuerdo al orden en que se encuentran, para cada variable.
3. Para cada par de números asignados, calcular la diferencia (d) entre el valor de sus datos y calcular el cuadrado de d.
4. Calcular la suma de los cuadrados de las diferencias.
5. Calcular la correlación de Spearman (p) con la siguiente formula:

$$\rho = 1 - \frac{\sum d^2}{N*(N^2-1)} \tag{3.1}$$

   Donde:
   N es el número de valores por variable.

Los resultados de la correlación de Spearman son entre -1 y 1. Un resultado cercano a 1 indica una alta correlación entre variables y un valor cercano a 0 indica una baja correlación, lo que se interpreta como poca relación entre variables.

3.3. Minería de datos

Minería de datos es la exploración y el análisis de grandes cantidades de datos para descubrir principalmente patrones y reglas [34]. Para el propósito de esta tesis, se considera que el éxito de la minería de datos es permitir a una empresa mejorar la NCP a

través del análisis de tráfico de su red de datos. Sin embargo, cabe mencionar que las herramientas y técnicas de minería de datos descritas aquí son igualmente aplicables en otras áreas de investigación como: astronomía, medicina, procesos industriales, mercadotecnia y economía.

La minería de datos de aplicaciones comerciales se ha basado en las investigaciones realizadas en estadística, ciencias de la computación y máquinas de aprendizaje. La elección de una combinación en particular de técnicas para aplicarlas a un problema en especial depende de la naturaleza de la tarea de minería de datos, el tipo de datos que se tengan disponibles y de la habilidad y preferencia de las personas que realicen la minería [45].

La minería de datos viene en dos sentidos: *directa e indirecta*. minería de datos directa permite explicar o categorizar algún campo en particular como entrada o respuesta, por ejemplo, la predicción de la variable utilización de los enlaces en una red de datos, donde la entrada es un conjunto de variables dependientes y la salida la predicción de la utilización del enlace. Minería de datos indirecta permite encontrar patrones o similitudes entre grupos de registros sin el uso de un campo en particular o colección de clases predefinidas [34]. Esta tesis esta enfocada a la minería de datos directa.

La minería de datos está concentrada en la construcción de modelos. Un modelo es simplemente un algoritmo o conjunto de reglas que conectan una colección de entradas a un target o salida en particular [34]. Regresión, RN, árboles de decisión son técnicas de minería de datos que permiten crear modelos [30], [20]. Un ejemplo sería un modelo de clasificación basado en árboles de decisión, el cual puede clasificar un conjunto de datos de acuerdo a un target en particular.

Las técnicas de minería de datos se pueden clasificar de acuerdo al tipo de tareas o problemas a los que se pueden aplicar [34]. Algunas de ellas son:

- Clasificación
- Estimación
- Predicción
- Afinidad de grupos
- Clustering
- Descripción

Las primeras tres tareas son ejemplos de minería de datos directa, donde el objetivo es encontrar el valor de una variable en particular. Afinidad de grupos y clustering son tareas indirectas donde el objetivo es descubrir estructuras en los datos sin emplear alguna variable como target.

Clasificación

Clasificación es una de las tareas más comunes en minería de datos. Para entender y comunicar el mundo, estamos constantemente clasificando y categorizando.

La clasificación consiste en examinar las características de un objeto y asignarlo dentro de un conjunto predefinido de clases [7]. Los objetos a ser clasificados son generalmente representados por registros en una tabla de un archivo o base de datos, y la clasificación consiste en agregar una nueva columna con un código clase de algún tipo. La clasificación se caracteriza por una definición de clases y un conjunto de entrenamiento que consiste de ejemplos preclasificados. La tarea es construir un modelo de algún tipo que pueda ser aplicado para clasificar datos no clasificados.

Ejemplos de las tareas de clasificación:

- Clasificar aspirantes a un crédito con un grado de confiabilidad bajo, medio o alto.
- Determinar qué tipo de alarmas generadas en una red de telecomunicaciones corresponden a un dispositivo de hardware en particular.

En todos estos ejemplos, hay un límite en el número de clases y en donde se espera asignar cada uno de los registros en alguna de las clases. Algunas técnicas de minería de datos que llevan acabo la clasificación son los *árboles de decisión* [14] y las *RN* [51].

Un árbol de decisión es una estructura de árbol donde cada nodo interno denota una prueba a un atributo, cada rama representa un resultado de la prueba, y los nodos hoja representan clases o distribución de clases [24]. Existen diversos criterios de decisión para seleccionar el atributo durante la construcción del árbol de decisión: Entropía (Quinlan, 1986) usado en el algoritmo C4.5, Gini-index (Breiman, 1984) usado en el algoritmo CART y Chi-square [35]. La ventaja de los árboles de decisión sobre las RN es que presentan mejor interpretación en los resultados.

RN se introduce en la sección 3.4.

Estimación

La clasificación se enfrenta con resultados discretos: si o no, nivel de riesgo bajo, medio o alto. La estimación se ocupa de salidas con valores continuos [9]. Dadas algunas entradas, la estimación produce un valor para alguna variable continua desconocida, por ejemplo, utilidades de una empresa o la altura de una persona.

En la práctica, la estimación se utiliza para desempeñar una tarea de clasificación. En una empresa de telecomunicaciones se podría construir un modelo de clasificación para saber si un enlace de datos en la red de telecomunicaciones ha sobrepasado un umbral indicando que el canal está sobrecargado, o desarrollar un modelo que no solo clasifique de esta manera sino que además nos pueda indicar el porcentaje de utilización del enlace.

La estimación tiene la gran ventaja que los registros individuales se pueden ordenar de acuerdo a su nivel de estimación. Para ver la importancia de esto imaginemos que una empresa que brinda servicios telefónicos pretende enviar 500,000 anuncios de publicidad a sus clientes. Si se utiliza la clasificación y se tienen 1.5 millones de clientes que regularmente responden a la publicidad, entonces podría simplemente enviar la publicidad a 500,000 clientes seleccionados de manera aleatoria. Por otro lado, si cada cliente tuviera un nivel de respuesta a la publicidad que se le envía, se podría enviar la publicidad a los 500,000 con mayor nivel de respuesta.

Ejemplos de las tareas de estimación son:

- Estimar el número de hijos en una familia;
- Estimar el nivel de ingresos de una familia;
- Estimar el tiempo que una persona será cliente de alguna empresa;
- Estimar la probabilidad que una persona responda a una publicidad.

Modelos de regresión y RN son técnicas de minería de datos que tratan el problema de la estimación.

Predicción

La predicción es lo mismo que clasificación y estimación, excepto que los registros están clasificados de acuerdo algún valor o comportamiento futuro. En una tarea de predicción la única forma de verificar la precisión de la clasificación es esperar y ver.

La razón principal por la que se considera la tarea de predicción por separado de la clasificación y la estimación es que en un modelo predictivo hay problemas adicionales con respecto a las relaciones temporales de las variables de entrada y la variable que se quiere predecir (target) [45].

Cualquiera de las técnicas usadas para clasificación y estimación se pueden adaptar para usarse en la predicción [40].

Ejemplos de las tareas de predicción que se pueden resolver con técnicas de minería de datos son:

- Predecir cuales de los clientes que tiene una empresa de telefonía ordenarán el nuevo servicio de correo por celular.
- Predecir variables económicas financieras [18].
- Predecir el precio del café de algún país [33].
- Predecir la carga eléctrica producida por alguna ciudad [15].

Elegir la técnica adecuada para la predicción depende de la naturaleza de los datos de entrada, el tipo de valores a predecir, y el nivel de interpretación que queremos en la predicción.

Afinidad de grupos

La tarea de afinidad de grupos es determinar qué cosas van juntas [34]. Un ejemplo sería determinar que cosas van juntas en un carro de compras en el supermercado. La afinidad de grupos se puede utilizar para planear el arreglo de productos en los almacenes de una tienda o en un catálogo con la finalidad de que objetos que regularmente se compran juntos se vean juntos.

Afinidad de grupos es una de las más simples aproximaciones para generar reglas de datos. Un ejemplo en el área de telecomunicaciones, sería el orden en que se generan alarmas dentro de la red de telecomunicaciones para detectar fallas en dispositivos de hardware, si dos alarmas, una alarma A generada por un dispositivo 1 se genera después de 1 minuto de una alarma B generada en un dispositivo 2, indica que el dispositivo 2 estará sobrecargado en 5 minutos, entonces esta frecuencia de alarmas nos permite generar asociación de reglas [25]:

Si la alarma A generada por el dispositivo 1 se genera después de 1 minuto de una alarma B generada en un dispositivo 2, con probabilidad P, entonces el dispositivo 2 estará sobrecargado.

Clustering

Clustering es la tarea de segmentar una población heterogénea dentro de un número de subgrupos más homogéneos o clusters [12]. Lo que distingue a clustering de clasificación es que en clustering no se tiene un grupo predefinido de clases. En clasificación, cada registro se asigna a una clase predefinida basada en un modelo desarrollado a través del entrenamiento de muestras preclasificadas.

En clustering no hay clases predefinidas y tampoco muestras con las que se haya entrenado el modelo. Los registros son agrupados de acuerdo a su propia similitud.

Clustering frecuentemente se utiliza para procesar datos que posteriormente servirán como entrada para algún otro modelo de minería de datos. Por ejemplo, clustering podría ser el primer paso en un proceso de segmentación de mercados: en lugar de preguntarse cuáles clientes responderán mejor a algún producto, primero se dividen los clientes en clusters o personas con determinados hábitos de compra, y entonces se identifica la mejor promoción para cada uno de los clusters.

Perfil y Descripción

En algunos casos el propósito de la minería de datos es simplemente describir que pasa en un conjunto de datos (una base de datos complicada), de tal forma que se incremente su entendimiento [34].

Como se mencionó anteriormente, la selección de cuál técnica de minería de datos se debe utilizar depende del tipo de problema que se quiera resolver. Se han desarrollado trabajos en el área de telecomunicaciones en donde se han resuelto problemas desde como una compañía telefónica puede utilizar un modelo de clustering para analizar los datos de sus clientes y poder hacer un estudio de mercado, hasta el analizar las alarmas generadas entre dispositivos de la red por medio de asociación de reglas para evitar fallas importantes en dispositivos de hardware [25].

En esta tesis se da una solución al problema de la predicción de tráfico en una red

de datos desarrollando un modelo predictivo basado en RN. Se utilizan RN gracias a su eficiente adaptación y buena capacidad de predicción en redes de datos [6].

En la siguiente sección se presentan los conceptos relacionados con RN: topologías, estructura, entrenamiento y aplicaciones.

3.4. Redes neuronales (RN)

Las RN son populares porque han demostrado ser útiles en muchas aplicaciones de minería de datos y de soporte de decisiones. Además de ser una técnica aplicada a predicción, pronóstico, clasificación y clustering. En el área de las telecomunicaciones son aplicables para la detección de fraudes, marketing, predicción de fallas en el equipo de telecomunicaciones, predicción de tráfico en redes de telecomunicaciones y pronóstico de variables de una red de datos.

Las RN tienen la habilidad de aprender en muchos aspectos como una persona experta en un área de trabajo tienen de su experiencia. Por ejemplo una persona que administra una red de datos y es experta en analizar los niveles de tráfico entre los enlaces de la red, podría saber por su experiencia que hay una gran carga de tráfico todos los días por la mañana debido a que en ese momento los usuarios de la red de datos, hacen su autenticación, consultan su correo y páginas de internet, además que hacen consultas a la base de datos para descargar información. De la misma manera que la persona ha aprendido con su experiencia es posible entrenar a una red neuronal con el historial de datos para que nos indique el momento en el que un enlace en la red esta sobrecargado.

La red neuronal aprendería de la misma forma que la persona experta. Necesita como entrada las mismas variables que la persona necesita para identificar la carga de tráfico, y como salida daría el nivel de sobrecarga del enlace, ver figura 3.1.

Las RN son ideales para problemas de predicción o estimación. Este tipo de problemas comparten las siguientes características:

- Las entradas son bien conocidas y se entienden claramente. Se tiene una idea de cuáles características de los datos son importantes, pero no se sabe como se combinan o si hay alguna relación entre ellas.
- Se conoce el tipo de salida que se quiere. Se conoce que es lo que se esta tratando de modelar.

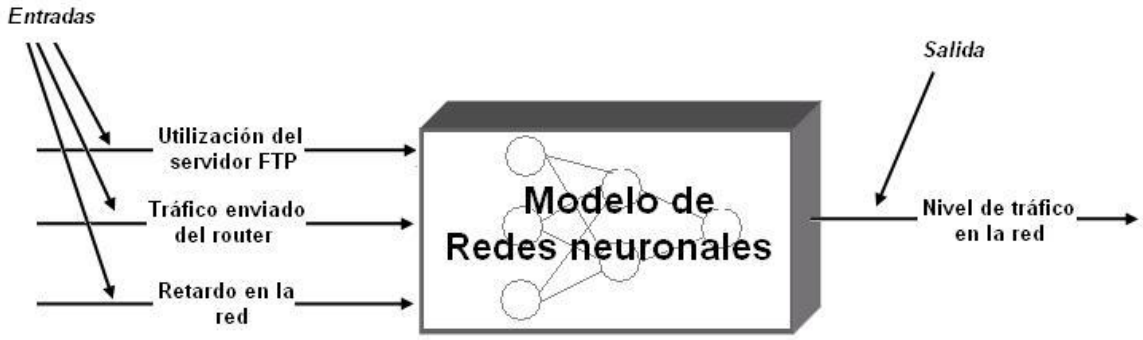

Figura 3.1.: La red neuronal es un proceso que sabe cómo procesar un conjunto de entradas para crear una salida.

- La experiencia está disponible. Se tienen gran cantidad de datos donde tanto las entradas como las salidas se conocen. Este conjunto de datos se utiliza para entrenar la red.

La manera en que una red neuronal se puede utilizar para resolver problemas de clasificación o predicción es la siguiente:

- Identificar las variables de entrada y de salida.
- Transformar las entradas y salidas a un formato adecuado para la red neuronal.
- Desarrollar una red neuronal con una topología apropiada.
- Entrenar la red con un conjunto de datos de entrenamiento.
- Usar un conjunto de datos de simulación para seleccionar los pesos que minimicen el error.
- Evaluar el desempeño de la red usando datos de prueba.
- Aplicar el modelo generado por la red para predecir una salida teniendo como entradas valores desconocidos.

Afortunadamente existe software de minería de datos que lleva acabo muchos de estos pasos de manera automática. Aunque no es necesario un conocimiento interno de la manera en que se implementa la red neuronal, hay algunas características que pueden mejorar su desempeño. Para modelos de predicción, el problema más importante es elegir los datos de entrenamiento correctos. Segundo, representar los datos de tal forma

que maximicen la habilidad de la red para reconocer patrones. Y tercero, interpretar los resultados de la red. Además de entender algunos detalles de cómo trabaja la red neuronal como la topología de la red y los parámetros que controlan el entrenamiento, pueden ayudar a mejorar el desempeño.

3.4.1. ¿Qué es una red neuronal?

Una Red Neuronal consiste de unidades básicas que simulan el comportamiento de neuronas biológicas encontradas en la naturaleza [34]. Por ejemplo hay una unidad dentro del sistema visual de los humanos que responde al movimiento de los objetos, y hay otra unidad que responde a los sonidos generados por los objetos. Estas unidades se conectan a una neurona que da una respuesta cuando el valor combinado de estas dos entradas es alto. Esta neurona es una entrada de otra que hace que la vista siga el movimiento de un objeto.

La idea básica es que una unidad neuronal, ya sea en un humano o una computadora, tenga varias entradas que la unidad combina en un solo valor de salida. En las computadoras, las unidades son conectadas simplemente unas con otras, como se muestra en las figuras 3.2, 3.3, 3.4, y 3.5, donde las salidas de algunas unidades se usan como entradas en otras.

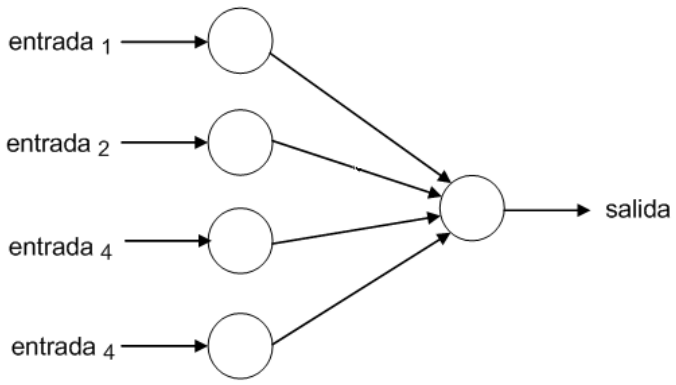

Figura 3.2.: Una RN que tiene cuatro entradas y produce una salida. El resultado de entrenar a esta red es equivalente a una técnica estadística llamada regresión logística.

La figura 3.6, muestra las características de una neurona artificial. La unidad combina sus entradas en un solo valor, el cuál se transforma para producir la salida, a estas

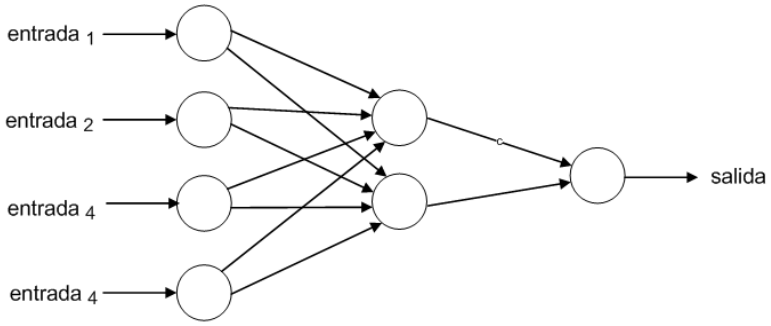

Figura 3.3.: Esta red tiene una capa intermedia llamada capa oculta, la cual hace a la red pueda reconocer más patrones.

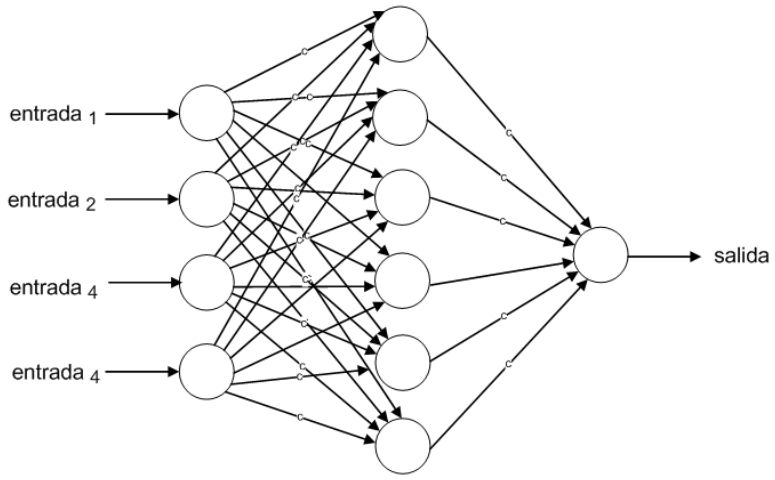

Figura 3.4.: Incrementar el número de unidades de la capa oculta hace que la red aumente su precisión pero introduce el riesgo de sobre entrenamiento. Usualmente solo se necesita una capa oculta.

dos acciones se les conoce como función de activación. Las funciones de activación más comunes están basadas en el modelo biológico, donde la salida permanece en un nivel bajo hasta que una combinación de las entradas alcanza un valor de umbral. Cuando la combinación de las entradas alcanza el umbral, la unidad se activa y la salida pasa a un nivel alto.

Una unidad en una red neuronal tiene la propiedad que pequeños cambios en las entradas, cuando los valores combinados están dentro de algún rango medio, puede tener grandes efectos en la salida y de manera inversa si hay grandes cambios en las

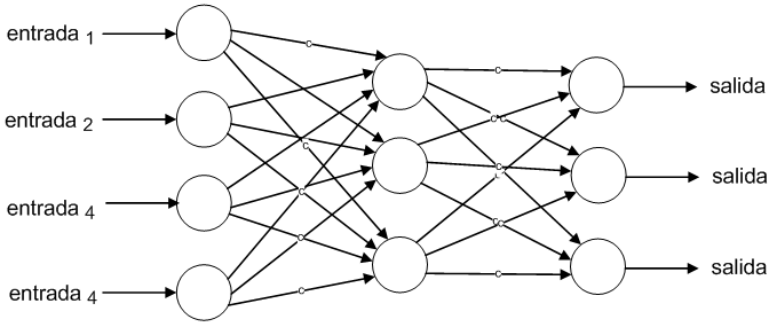

Figura 3.5.: Una red neuronal puede producir múltiples valores de salida.

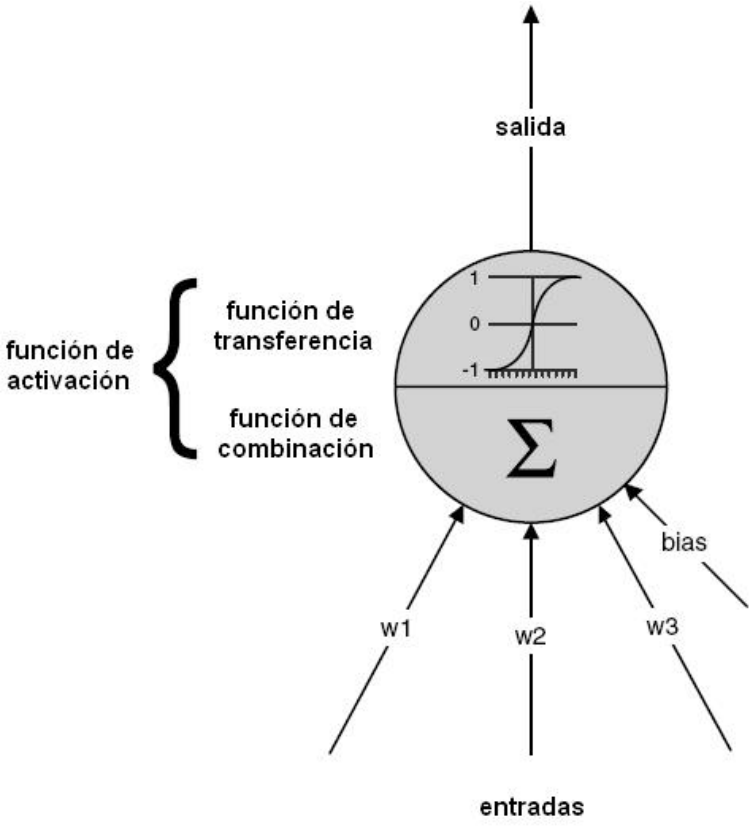

Figura 3.6.: La unidad de una red neuronal artificial modela una neurona biológica. La salida de la unidad es una combinación no lineal de sus entradas.

entradas pueden tener pequeños efectos en la salida, cuando la combinación de las en-

tradas están lejos del rango medio. Esta propiedad, en la que algunas veces pequeños cambios importan y algunas veces no, es un ejemplo de un comportamiento no lineal. La potencia y complejidad de las RN surge de su comportamiento no lineal, el cual se logra de acuerdo al tipo de función de activación usada por la unidad de la neurona.

La función de activación tiene dos partes. La primer parte es la función de combinación la cual une todas las entradas en un solo valor. Como se muestra en la figura 3.6, cada entrada en la unidad tiene asociado un peso. La función de combinación más común es la suma de los pesos, donde cada entrada es multiplicada por su propio peso y todos sus productos se suman. La elección de la función de combinación son algunas de las características de las RN.

La segunda parte de la función de activación es la función de transferencia, su nombre se origina del hecho de transferir el valor de la función de combinación a la salida de la unidad. La figura 3.7, compara tres funciones de transferencia: la función sigmoid (logística), lineal y tangente hiperbólica. Una red neuronal "feed-forward" consiste solo de unidades con funciones de transferencia lineales y una función de combinación que suma los pesos y que solo hace una regresión lineal. Las funciones sigmoid son funciones S-formadas, de las cuales las dos más comunes para RN son la logística y la tangente hiperbólica. La diferencia mas grande es el rango de sus salidas, entre 0 y 1 para la logística y entre -1 y 1 para la tangente hiperbólica.

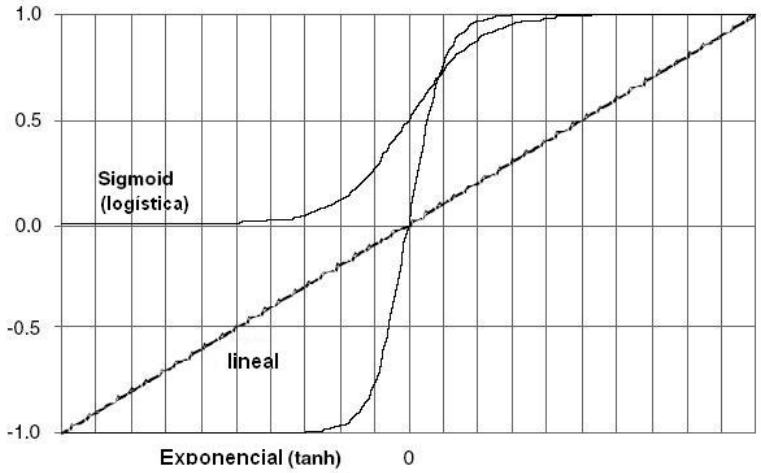

Figura 3.7.: Tres funciones de transferencia comunes son: sigmoid, lineal y la tangente hiperbólica.

Las formulas de estas funciones son:

$$logistic(x) = 1/(1 + e^{-x}) \tag{3.2}$$

$$tanh(x) = (e^{x} - e^{-x})/(e^{x} + e^{-x}) \tag{3.3}$$

Para estas funciones cuando la suma de los pesos de todas las entradas es cercana a cero, entonces las funciones tienen una aproximación cercana a una función lineal. Conforme la magnitud de la suma de los pesos se incrementa, estas funciones de transferencia gradualmente se saturan (a 0 o 1 en el caso de la logística; a -1 o 1 en el caso de la tangente hiperbólica). Este comportamiento corresponde a un movimiento gradual de un modelo lineal de la entrada a un modelo no lineal en la salida. En resumen, las RN tienen la habilidad de hacer un buen trabajo en el modelo de tres tipos de problemas: problemas lineales, problemas cercanos a la linealidad y problemas no lineales.

3.4.2. ¿Como aprende una red neuronal?

Entrenar una red neuronal es el proceso de encontrar los mejores pesos en los arcos que conectan todas las unidades en la red. El éxito es usar un conjunto de datos de entrenamiento para calcular los pesos de tal manera que la salida de la red sea un valor cercano al valor deseado. El algoritmo de *retro-propagación* [28], fue el algoritmo original para entrenar redes multicapa. Los tres principales pasos en el algoritmo de retro-propagación son:

- La red entrena un ejemplo, usa los pesos existentes en la red y calcula la salida o salidas.
- El algoritmo de retro-propagación entonces calcula el error tomando la diferencia entre el resultado calculado y el esperado (resultado actual).
- El error se retro-alimenta a la red y los pesos son ajustados para minimizar el error, de aquí el nombre de retro-propagación porque el error se envía hacia atrás de la red.

El algoritmo de retro-propagación mide el error total de la red [28].

3.4.3. Topologías

La topología o estructura de una red neuronal con "feed-forward" es típica en redes usadas para predicción o clasificación. Las unidades son organizadas en tres capas. La primera capa se conecta a las entradas y se llama capa de entrada. La siguiente capa se llama capa oculta, cada unidad de esta capa regularmente se conecta a todas las unidades de la capa de entrada. Las unidades de la capa oculta calculan su salida multiplicando el valor de cada entrada con su correspondiente peso, sumando todos estos productos y aplicando la función de transferencia. Una red neuronal puede tener cualquier número de capas ocultas, pero en general una capa es suficiente [1]. La última capa es la capa de salida la cuál se conecta a la salida de la red neuronal. Esta capa esta conectada a todas las unidades de la capa oculta. En muchas ocasiones la red neuronal se utiliza para calcular solo una salida, en este caso la capa de salida solo tiene una unidad.

Perceptron

Una de las primeras arquitecturas de RN fue Perceptron [51], el cual es un tipo de modelo lineal. Perceptron usa una combinación lineal como función de combinación. En la práctica su función de activación es casi siempre una función logística, lo cuál hace que Perceptron se comporte como un modelo de regresión logístico. Un ejemplo de una arquitectura Perceptron con 4 variables de entrada y una variable de salida se muestra en la figura 3.2.

Perceptron Multicapa (MLP)

Las RN pueden tener transformaciones por medio de una capa oculta. Generalmente, cada unidad de entrada se conecta con cada unidad en la capa oculta y cada unidad oculta se conecta a cada unidad de salida. Las unidades de la capa oculta combinan los valores de entrada y aplican una función de activación, que puede ser no-lineal. Los valores calculados por las unidades de la capa oculta se combinan en las unidades de salida. Es común tener dos o tres capas ocultas, donde cada capa oculta se conecta a la siguiente capa oculta, hasta que la última capa oculta se conecta a la capa de salida. Si este tipo de red neuronal usa una función de combinación lineal y una función de transferencia sigmoid se le llama Perceptron Multicapa (MLP) [41]. La figura 3.4 muestra una red neuronal MLP con una capa oculta, 4 variables de entrada y una variable de salida.

3.4.4. Modelo para predicción

Como se menciono anteriormente hay muchos modelos para solucionar varias de las tareas de la minería de datos. Estos modelos varían en varios aspectos, como: el tiempo que requieren para su aprendizaje, tolerancia al ruido en los datos, el formato esperado de los datos y los conceptos que son capaces de expresar.

Las técnicas de minería de datos se han aplicado a tareas de clasificación. Se han investigado varias técnicas de clasificación durante los últimos 20 años, el resultado han sido varios algoritmos que encuentran patrones predictivos en conjuntos de datos [45]. Los datos usados en la tarea de clasificación consisten en un conjunto de casos, donde cada uno de ellos está representado por un conjunto de variables independientes y una variable dependiente. La tarea de predicción consiste en predecir la variable dependiente basándose en las variables independientes [45].

La tabla 3.1 muestra la representación de los datos para la tarea de predicción. Cada fila representa un caso. Para cada caso, las columnas representan los valores de las n variables independientes. La última columna es especial porque representa la variable dependiente del caso, es decir el valor que la tarea de predicción trata de predecir.

Tabla 3.1.: Formato de los datos de entrenamiento del modelo de predicción.

Casos	v_1	v_2	v_3	.	v_{n-2}	v_{n-1}	v_n	target
1	$v_{1,1}$	$v_{2,1}$	$v_{3,1}$	.	$v_{n-2,1}$	$v_{n-1,1}$	$v_{n,1}$	t_1
2	$v_{1,2}$	$v_{2,2}$	$v_{3,2}$	.	$v_{n-2,2}$	$v_{n-1,2}$	$v_{n,2}$	t_2
3	$v_{1,3}$	$v_{2,3}$	$v_{3,3}$	.	$v_{n-2,3}$	$v_{n-1,3}$	$v_{n,3}$	t_3
4	$v_{1,4}$	$v_{2,4}$	$v_{3,4}$	.	$v_{n-2,4}$	$v_{n-1,4}$	$v_{n,4}$	t_4
5	$v_{1,5}$	$v_{2,5}$	$v_{3,5}$	.	$v_{n-2,5}$	$v_{n-1,5}$	$v_{n,5}$	t_5

Un modelo de predicción se puede desarrollar con una Red Neuronal MLP con tres capas (entrada, oculta y salida). Donde el número de entradas de la red depende del número de variables independientes. La ventaja de este modelo de RN entrenado por un algoritmo de retro-propagación es que puede aproximar cualquier función no lineal [22].

3.4.5. Modelo para pronóstico

La investigación en RN se ha concentrado en pronosticar el desarrollo de series de tiempo. Formalmente se podría decir que una red neuronal encuentra una función f: $\Re^N \rightarrow \Re$ para poder obtener un valor de una serie de tiempo x en el tiempo t+d, basándose en los N tiempos anteriores de t, de tal forma que:

$$x(t+d) = f(x(t), x(t-1), ..., x(t-N+1)) \tag{3.4}$$

Usualmente el valor de d es uno, para que f pronostique el siguiente valor de x.

Las RN se han utilizado en varias aplicaciones para el pronóstico de series de tiempo, algunas de ellas son: pronóstico de la carga de electricidad [37] y pronóstico del desempeño de una red [16].

La figura 3.8 da la idea básica del pronóstico con RN.

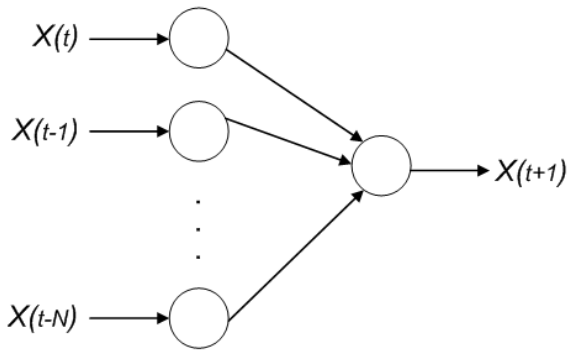

Figura 3.8.: Arquitectura de una red neuronal perceptron para el pronóstico de series de tiempo.

Una Red Neuronal Perceptron con la arquitectura que se muestra en la figura 3.8 representa un modelo de regresión que se puede utilizar para desarrollar un modelo de Vector Autoregresión (VAR).

El uso de modelos VAR se ha recomendado por Sims [11], como una alternativa eficiente para verificar relaciones causales en variables económicas y pronosticar su evolución. Dado el vector de variables y_t con k variables, el modelo VAR explica cada variable con sus propios p valores anteriores y los p valores anteriores del resto de las variables, de acuerdo con la relación:

$$y_t = d_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t \tag{3.5}$$

$$y_t = d_0 + \sum_{j=1}^{p} \phi_j \, y_{t-j} + \varepsilon_t \tag{3.6}$$

donde:
d_0 es un vector de (k x 1),
ϕ_j es una matriz (k x k) de coeficientes autoregresivos (j=1,2,...,p),
ε_t es un vector de (k x 1) con componentes de ruido blanco.

VAR se ha usado en aplicaciones económicas y financieras para tomar decisiones en política económica y en el estudio de series de tiempo de datos para mantener la coherencia de variables económicas [8], [13], [50].

Los datos históricos de la variable que se quiere pronosticar se utilizan para construir un modelo que explique el comportamiento actual de la variable, de esta manera cuando el modelo se aplique a valores actuales, el resultado será un pronóstico del comportamiento futuro de la variable.

3.5. Discusión

En este capítulo se mencionaron las tareas que se pueden resolver con inteligencia analítica. Además, se explicó una técnica de minería de datos, como la RN que se puede utilizar para resolver el problema de pronóstico de series de tiempo y la predicción de variables. Se mencionaron las diferentes topologías de una RN, y como su estructura y función de activación que utilice definen los diferentes tipos de RN. Además se menciono como una red neuronal perceptron gracias a sus características de regresión lineal se puede utilizar como un modelo VAR, para el pronóstico de variables.

Este capítulo sirve como marco teórico de los conceptos que se utilizarán en una metodología que se propone para resolver el problema de predicción de tráfico en enlaces de redes de datos.

Capítulo 4.

Metodología desarrollada

Como mencionamos anteriormente debido a la gran cantidad de estadísticas que se generan en una red de datos sería útil analizarlas para ayudar a resolver los problemas que están involucrados con la NCP. La predicción de tráfico es uno de los problemas que se pueden resolver. Debido a la naturaleza distribuida de los dispositivos de la red y la variedad de servicios que se ofrecen es complicado seleccionar y pronosticar aquellas variables que estén relacionadas con los niveles de utilización de los enlaces de la red de datos. En este trabajo de investigación planteamos una metodología en la que se utilizan técnicas de Inteligencia Analítica para resolver este problema.

4.1. Proceso de adquisición de conocimiento

La minería de datos hace fácil la tarea de aplicar algoritmos, tales como redes neuronales, árboles de decisión y algoritmos genéticos, a la gran cantidad de datos generados en el área de telecomunicaciones para poder descubrir conocimiento que era desconocido. Sin embargo, a pesar de que los algoritmos son importantes, un proceso de adquisición de conocimiento es más que solo una estructura de datos y técnicas poderosas. Las técnicas tienen que ser aplicadas en las áreas correctas en los datos correctos. El círculo virtuoso del proceso de adquisición de conocimiento es un proceso de aprendizaje iterativo que da soluciones con el tiempo [34].

La figura 4.1 muestra un círculo virtuoso de un proceso de adquisición de conocimiento en un ambiente de telecomunicaciones, cuyas etapas son cuatro:

1. *Identificar el problema.* Consiste en identificar las áreas donde el análisis de datos puede tener algún valor.

   En un ambiente de telecomunicaciones este paso consiste en analizar los diferentes datos generados en la red (datos de los clientes, datos de las llamadas o datos de

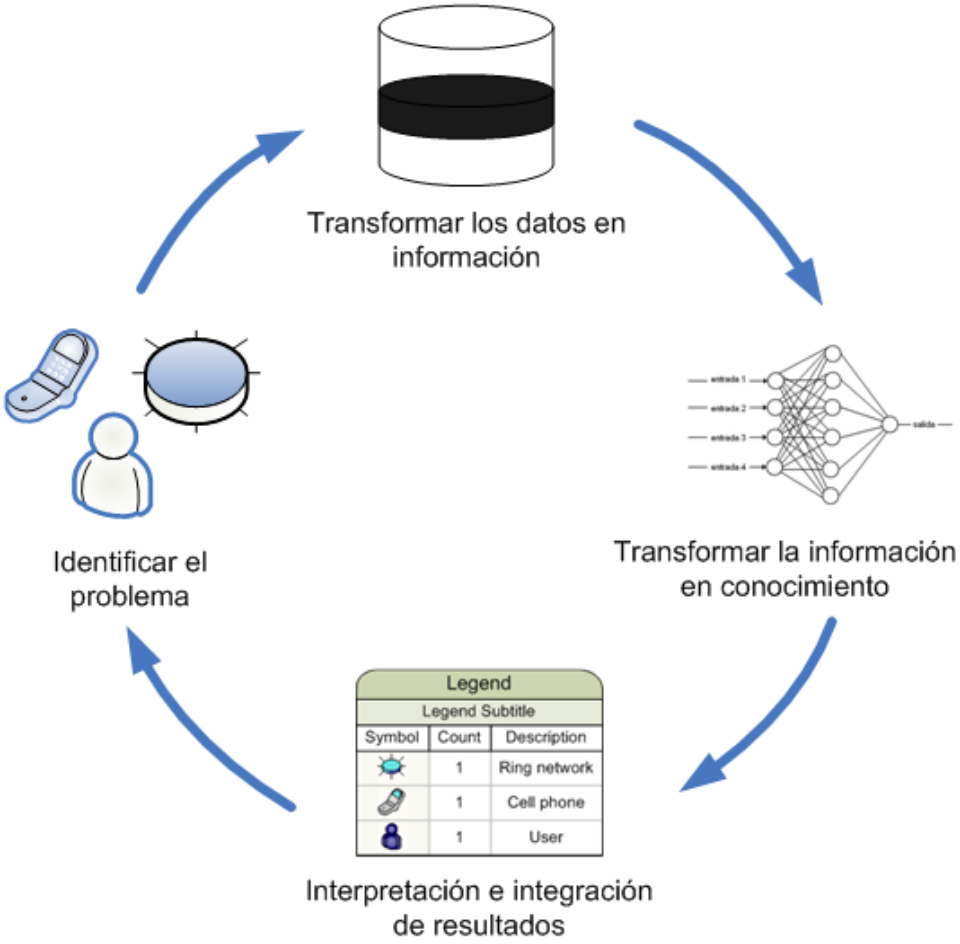

Figura 4.1.: Proceso de adquisición de conocimiento en un ambiente de telecomunicaciones

la red), y enfocarlos de acuerdo al problema que se quiera resolver. Algunos de estos problemas son [48]:

- Analizar las estadísticas de los dispositivos de hardware de la red para poder resolver el problema de predicción de tráfico.
- Analizar la frecuencia de las llamadas de los usuarios de una empresa telefónica para la detección de llamadas fraudulentas.
- Utilizar los datos de los clientes para hacer un estudio de mercado o detectar perfiles de los clientes.

Estas estadísticas se pueden obtener de las siguientes fuentes de datos:

- De los dispositivos de una red de datos del mundo real.
- De un simulador de redes de telecomunicaciones como Opnet Modeler [3] ver anexo A, que brinda un ambiente para diseñar y estudiar dispositivos, protocolos y aplicaciones de redes de telecomunicaciones.

Estas estadísticas de red se pueden analizar para llevar acabo la predicción de tráfico en los enlaces de una red de datos. Sin embargo tomar estas estadísticas como entradas de manera directa de un modelo de predicción podría originar un error muy grande en los resultados, o incluso no poder llevar acabo ninguna predicción. Para resolver este problema es necesario obtener información a partir de estos datos y darles la representación más adecuada de acuerdo al modelo de predicción que se vaya a utilizar.

2. *Transformar los datos en información.* Consiste en utilizar técnicas analíticas para convertir los datos en información. Esta información consiste en encontrar datos de los datos, es decir analizar los datos para encontrar características que mejor los representen. Estas técnicas se pueden clasificar de acuerdo al procesamiento que hacen sobre los datos:

   - Limpieza de datos. Se puede aplicar técnicas para remover ruido, corregir inconsistencias en los datos y detectar valores de variables fuera de rango. Por ejemplo, si una variable tiene valores infinitos se podrían sustituir por el valor promedio de sus datos.
   - Integración de datos. Resuelve el problema de redundancia en la integración de datos de múltiples fuentes como base de datos de alarmas generadas en una red de telecomunicaciones, estadísticas de utilización de dispositivos de hardware, archivos exportados de un simulador de redes de telecomunicaciones, etc. Por ejemplo la variable utilización de un enlace de datos en una red, es redundante ya que se deriva de la variable throughput. Un análisis de correlación nos permite resolver este problema, algunas técnicas de correlación son las de Pearson y Spearman.
   - Transformación de datos. Los datos se transforman en formas apropiadas para la minería de datos. Algunas técnicas de transformación de datos son: la normalización, smoothing, agregación, generalización y estandarización. Por ejemplo algunos modelos de regresión no lineal se pueden convertir a

modelos lineales aplicando transformaciones en las variables de entrada [34]. Otro ejemplo sería normalizar los valores de una variable en un rango de 0 y 1.

- Reducción de datos. Técnicas de reducción de datos buscan reducir la representación de los datos, pero manteniendo la integridad de los datos originales. Ejemplos de estas técnicas son: aplicar operaciones de agregación en los datos para la construcción de un cubo de datos y detección de variables irrelevantes por medio de algoritmos de árboles de decisión como C4.5 [47].

Una estadística de un dispositivo de hardware de la red de datos la entenderemos como una variable de red. La transformación de los datos consiste en aplicar técnicas analíticas a las m variables de red, para seleccionar aquellas que representen mejor a los datos. Las n variables seleccionadas representan la información encontrada de los datos, ver figura 4.2.

Debido al cambio constante del tráfico que pasa por los dispositivos de hardware, es muy probable que las estadísticas obtenidas presenten inconsistencia y ruido en sus valores, por ello es necesario analizar los datos y aplicar una técnica de limpieza para lograr su consistencia, y poder aplicarlas como entrada de un modelo de predicción

Además existe el problema del gran número de variables obtenidas de una red de telecomunicaciones. Es posible que dentro de esta gran cantidad de variables existan algunas que sean redundantes y otras que sean irrelevantes y no contribuyan con la predicción del tráfico. Se pueden aplicar técnicas de correlación y selección para resolver este problema. Generalmente el número n de variables de salida, es menor que el número m de variables de entrada.

Una vez seleccionadas y estandarizadas las variables de red se tiene que seleccionar un modelo de predicción.

3. *Transformar la información en conocimiento.* Esta transformación se realiza aplicando técnicas de minería de datos. Es decir aplicar técnicas como RN, árboles de decisión, clustering o algoritmos genéticos a un conjunto de datos de entrada para obtener patrones o relaciones entre ellos.

   La información se considera a las variables seleccionadas y estandarizadas, estas variables son entrada de un modelo de predicción él cuál llevará acabo la transformación de esta información en conocimiento. Esta transformación consiste en

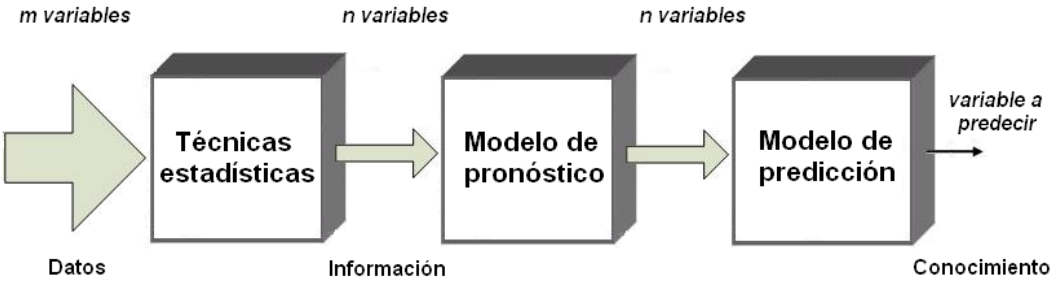

Figura 4.2.: Transformación de los datos a conocimiento.

predecir el valor de una variable dependiente basándose en los valores de un conjunto de variables independientes. El conocimiento obtenido será la predicción del tráfico, ver figura 4.2.

Como se explico en el capitulo de minería de datos, la tarea de predicción se puede resolver aplicando técnicas como árboles de decisión, RN o algoritmos genéticos.

4. *Transformar el conocimiento en inteligencia.* Consiste en la selección de patrones interesantes y la interpretación de los resultados para incorporar el conocimiento adquirido a un sistema experto[1].

   Esta etapa consiste en interpretar y entender el conocimiento, en muchas aplicaciones se refiere en incorporar el conocimiento adquirido a un sistema experto [45], para que pueda dar respuesta a un problema de manera automática. Esta aplicación del conocimiento adquirido es lo que se interpreta como inteligencia.

4.2. Predicción de tráfico en redes de telecomunicaciones

El proceso de adquisición de conocimiento como se mencionó en la sección anterior consiste en un círculo que involucra cuatro etapas. Sin embargo en cada una de estas etapas intervienen varios pasos. Identificar un problema envuelve realmente todo un proceso de obtención de fuentes de datos. La transformación de datos en información involucra varias técnicas analíticas. La transformación de información en conocimiento consiste en analizar de la gran variedad de algoritmos de predicción cual de ellos nos

[1]Un sistema experto es un programa que presenta y aplica conocimiento en alguna área de trabajo [45].

permite llevar acabo la predicción de acuerdo al tipo de datos que tenemos. La interpretación de resultados no solo consiste en una métrica de error, realmente consiste en verificar si los resultados corresponden a lo esperado. El propósito de esta sección consiste en establecer una metodología que nos permita analizar las estadísticas de una red de telecomunicaciones para resolver el problema de la predicción del tráfico en enlaces de redes de datos, basándose en técnicas de inteligencia analítica.

Los pasos a seguir en esta metodología no están enfocados solo en resolver el problema de predicción. Durante el análisis de predicción se van ir dando resultados parciales que pueden ser importantes para entender el comportamiento de la red de datos. Además, esta metodología de acuerdo al caso de estudio que abordaremos en el capítulo 5 estará enfocada a la predicción de una sola variable de red, sin embargo este conjunto de pasos que integran la metodología se podrían adaptar para la predicción de cualquier otra variable de red.

La metodología que describimos en esta sección consiste en los siguientes pasos:

1. *Colección de datos de un escenario de red:* Coleccionar estadísticas que describan el comportamiento de la red de datos.
2. *Preparación y limpieza de los datos.*
3. *Selección de variables no redundantes:* Análisis de correlación de las variables redundantes de la red.
4. *Selección de variables para la predicción.* Eliminar variables irrelevantes.
5. *Pronóstico de las variables seleccionadas que se utilizan en el modelo predictivo:* Pronosticar una serie de tiempo. La serie de tiempo son secuencias de las medidas de las variables de red seleccionadas.
6. *Predicción del target (tráfico de la red):* Predecir el target teniendo como entrada las variables pronosticadas.
7. *Interpretación y evaluación de los resultados:* Interpretar y evaluar los resultados predictivos producidos por el modelo de RN.

Como se muestra en la figura 4.3, la metodología no se desarrolla solo como un conjunto de pasos con algún orden, el proceso es recursivo ya que en muchas ocasiones si un paso en la metodología no cumple con los resultados esperados es necesario volver a etapas anteriores para hacer correcciones. Un ejemplo sería el momento de evaluar la

precisión de los modelos de predicción y si estos presentan un nivel de error muy grande, entonces se tiene que regresar a la etapa de construcción del modelo para modificar sus propiedades, o tal vez no sea problema del modelo, si no de las variables que se están seleccionando como entradas.

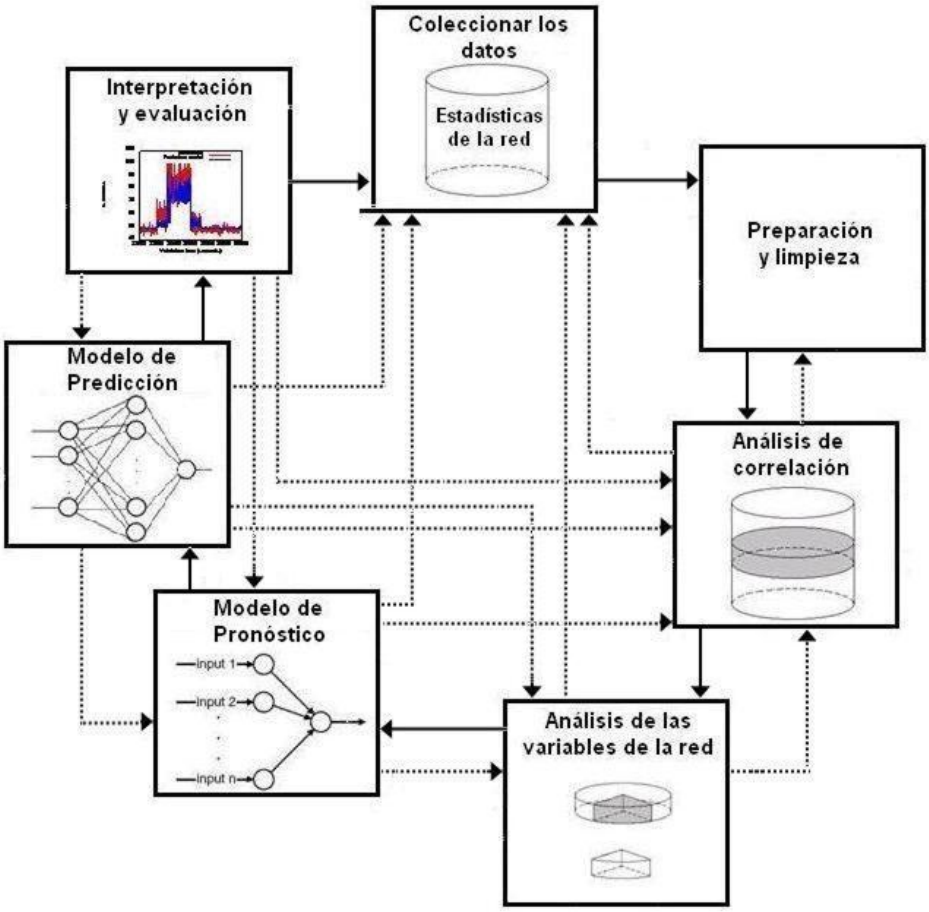

Figura 4.3.: Metodología para la predicción de tráfico.

4.2.1. Coleccionar los datos de un escenario de red

El objetivo de esta tesis es analizar las estadísticas generadas en dispositivos de hardware de una red de datos, para poder predecir los niveles de utilización de los enlaces de la red. Como se menciono en el capítulo 2 se considera como variable de red a cada

una de las estadísticas generadas por un dispositivo de red.

Para coleccionar estas estadísticas se utilizo la herramienta Opnet Modeler. Este simulador nos permite diseñar, construir y estudiar redes, dispositivos, protocolos y aplicaciones de comunicaciones, con una gran flexibilidad de poder variar las características de cada uno de ellos.

Cabe mencionar que esta metodología funciona también para estadísticas de escenarios de redes reales capturadas gracias a agentes.

Cada una de las estadísticas que se obtienen como resultado de la simulación de un escenario de red de datos en Opnet Modeler, puede representar una serie de tiempo, un ejemplo es como el nivel de utilización del CPU de un servidor de base de datos varia con el tiempo, como se muestra en la figura 4.4.

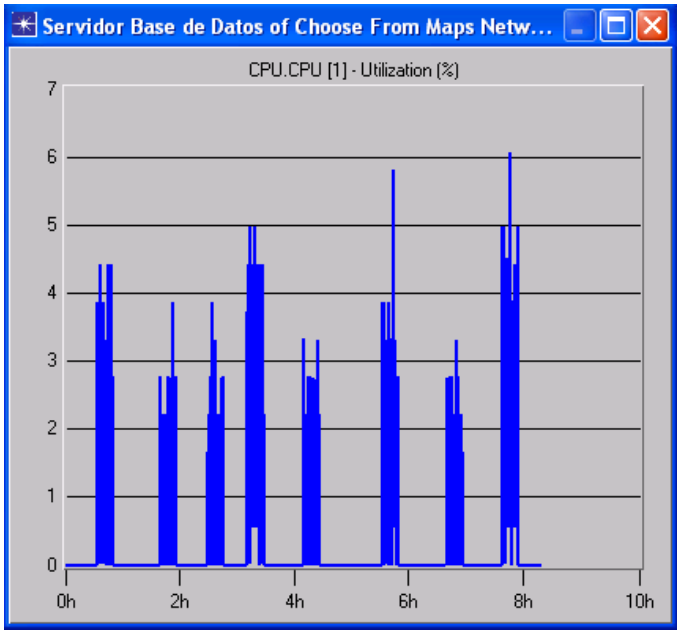

Figura 4.4.: Estadística de la utilización del CPU de un servidor de base de datos.

Opnet Modeler permite exportar las estadísticas generadas en la simulación a una hoja de cálculo, de esta manera se puede construir una serie de tiempo que tenga los valores de todas las variables de red.

La simulación en Opnet Modeler genera estadísticas que describen el comportamiento de una red de datos y permite analizarlas de manera individual, sin embargo seria conveniente conocer la dependencia y relación entre variables de red.

En el mundo real es difícil de obtener todas las estadísticas que se generan en una simulación, de aquí la importancia en poder conocer cuáles son las variables más importantes. Por este motivo surge la necesidad de crear una metodología que describa un proceso que permita adquirir conocimiento e inteligencia de la gran cantidad de estadísticas generadas en una red de datos.

4.2.2. Preparación y limpieza de los datos

El segundo paso de la metodología consiste en dar consistencia a los datos. El problema encontrado consistió en que en algunas muestras de la serie de tiempo no tenían valores definidos. Un ejemplo de ello se puede observar en la figura 4.5, donde se muestra que la estadística de la carga de tareas en el servidor de red no tiene un valor definido en el tiempo t = 120 minutos. Al exportar esta estadística no hay un valor definido para todos los puntos de la serie de tiempo.

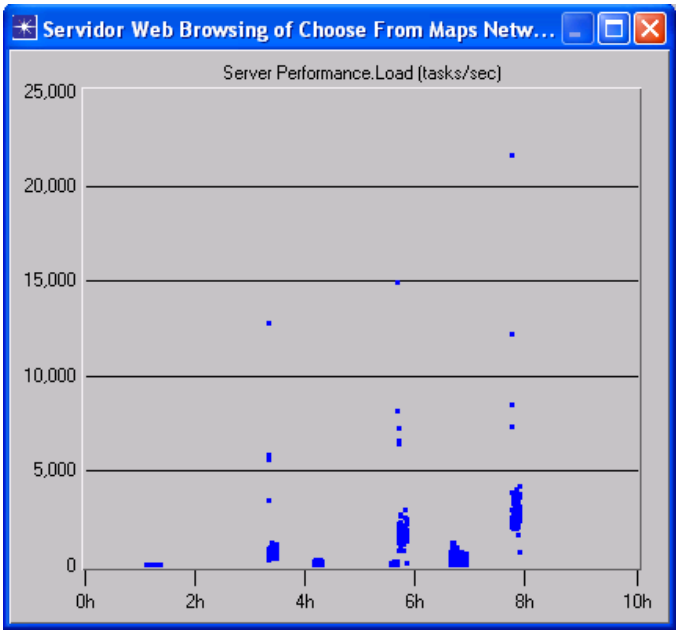

Figura 4.5.: Estadística de la carga de tareas por segundo del servidor web.

Para solucionar este problema se realizo un programa en Borland C, ver Apéndice B, en donde todas las muestras de la serie de tiempo que no tuvieran un valor definido se insertaba un valor constante igual a cero.

Con la eliminación de ruido en los datos, se puede obtener la siguiente información de las variables que integran la serie de tiempo:

- Identificar tipo de variables en la serie de tiempo: unarias, binarias, nominales, ordinales o de intervalo.
- Encontrar medidas estadísticas por cada variable: valor mínimo y máximo, media y desviación estándar.
- Representación gráfica de la distribución de las variables.
- Transformación de variables: crear variables a partir de otras. Por ejemplo crear una variable binaria a partir de una de intervalo.

Sin embargo un análisis estadístico solo permite obtener información de manera individual por cada una de las variables de la serie de tiempo, aun no se conoce la relación o dependencia entre variables.

4.2.3. Selección de variables no redundantes

Dos variables se consideran redundantes si ambas variables proporcionan la misma información, un ejemplo es la edad de una persona y su fecha de nacimiento, ambas variables representan la misma información. Identificar a las variables que puedan representar toda la información, en aplicaciones como base de datos puede ayudar a evitar espacio en disco. En las redes de datos identificar variables redundantes ayuda a delimitar el número de variables que se tienen que analizar para conocer el comportamiento de todo un sistema (una red de telecomunicaciones) y también para facilitar la captura de estas variables en escenarios reales.

Para llevar acabo la correlación de Spearman entre m variables de la serie de tiempo, se utilizo SAS Enterprise Miner, que genera una matriz de correlación de mx m valores.

La selección de las variables no redundantes se lleva acabo de la siguiente manera, ver :

- Establecer un umbral u para detectar una alta correlación entre variables.

- Si un valor de correlación $v_{i,j}$ entre las variables i y j, sobrepasan el umbral u, se considera una alta correlación, y solo se debe seleccionar alguna de las dos variables dentro del conjunto de variables no redundantes.
- Si un valor de correlación $v_{i,j}$ entre las variables i y j, no sobrepasan el umbral u, se considera una baja correlación, y ambas variables se incluyen dentro del conjunto de variables no redundantes.
- Si un valor de correlación $v_{k,j}$ entre las variables k y j, no sobrepasan el umbral u, antes de incluir las variables k y j dentro del conjunto de variables no redundantes, se debe verificar que estas no estén en el conjunto.

Este analísis de correlación se realizo en Borland C, ver Apéndice C. Como resultado del análisis de correlación se obtiene las variables no redundantes de la red. Ahora es conveniente saber la relación que hay entre estas variables y la variable que se quiere predecir.

4.2.4. Selección de variables para la predicción

Hay algunas variables que pueden definir el desempeño de la red tal como: *el nivel de utilización de los enlaces de red*, *carga de trabajo en los servidores*, *retardo en la red*, *tiempo de respuesta de las aplicaciones*, etc. Sin embargo, de acuerdo a la idea de este trabajo de investigación se seleccionó como variable a predecir a la variable que mide el nivel de utilización de los enlaces de red, debido a que se quiere conocer el momento en que un enlace de red es sobrecargado.

Dentro de las variables no redundantes pueden existir variables irrelevantes que no contribuyen a la predicción del target. Además un gran número de variables de entrada pueden afectar los modelos de predicción de varias maneras. Primero, entre más variables se utilicen como entrada se incrementa el riesgo de sobre entrenar el modelo, además de incrementar el tamaño de los datos de entrenamiento. Segundo, en caso de utilizar un modelo de predicción basado en RN, entre mayor sea el número de variables los pesos en la red neuronal tienden a ser menos óptimos. Este problema de selección es un problema común para estadísticas y los árboles de decisión son un buen método para elegir las mejores variables [34].

El objetivo de la selección de las variables no esta limitado solo en disminuir el número de variables de entrada del modelo de predicción. Realmente se busca conocer aquellas variables que nos sirvan como indicativo de los cambios en los niveles de tráfico de los enlaces de una red de datos. Un ejemplo sería que después del análisis de selección de

variables se detectará que la variable del tráfico enviado por el servidor de base de datos (bits/sec) contribuye de manera directa con la predicción del tráfico, de esta manera el identificar esta variable nos ayudaría a saber que el momento en el que el nivel de tráfico enviado por el servidor de base de datos se incremente provocará que el nivel de tráfico en algún enlace de red también lo haga.

Como resultado del análisis de selección de variables con respecto al target se obtienen las variables que son entrada del modelo de predicción.

4.2.5. Pronostico de las variables seleccionadas.

La principal tarea del modelo de predicción es tener n variables de entrada y una salida. La salida corresponde al valor que predice el modelo, ver figura 4.2.

De esta manera de acuerdo a los valores de las variables de entrada en el tiempo t se va a poder conocer el valor del target en el tiempo t. Sin embargo la tarea de la predicción es conocer el valor del target en un tiempo t+1, para ello se necesita pronosticar el valor de las variables de entrada en un tiempo t+1, como se muestra en la tabla 4.1, para un modelo de predicción con n entradas y un target de salida.

Tabla 4.1.: Formato de los datos de entrenamiento del modelo de predicción.

tiempo	v_1	v_2	v_3	.	v_{n-2}	v_{n-1}	v_n	target
t-4	$v_{1,t-4}$	$v_{2,t-4}$	$v_{3,t-4}$	.	$v_{n-2,t-4}$	$v_{n-1,t-4}$	$v_{n,t-4}$	$target_{t-4}$
t-3	$v_{1,t-3}$	$v_{2,t-3}$	$v_{3,t-3}$	.	$v_{n-2,t-3}$	$v_{n-1,t-3}$	$v_{n,t-3}$	$target_{t-3}$
t-2	$v_{1,t-2}$	$v_{2,t-2}$	$v_{3,t-2}$	.	$v_{n-2,t-2}$	$v_{n-1,t-2}$	$v_{n,t-2}$	$target_{t-2}$
t-1	$v_{1,t-1}$	$v_{2,t-1}$	$v_{3,t-1}$	.	$v_{n-2,t-1}$	$v_{n-1,t-1}$	$v_{n,t-1}$	$target_{t-1}$
t	$v_{1,t}$	$v_{2,t}$	$v_{3,t}$	.	$v_{n-2,t}$	$v_{n-1,t}$	$v_{n,t}$	$target_t$
t+1	-	-	-	.	-	-	-	X

Como se menciono en el capítulo 3, el pronostico de variables se puede realizar con un modelo de regresión lineal.

Un modelo VAR se puede implementar por medio de una RN tipo perceptron, como se explico en el capitulo 3. La figura 4.6 muestra la arquitectura Perceptron con n unidades de entrada y n variables de salida que se van a pronosticar.

La forma de entrenar una red neuronal Perceptron consiste en tener como datos de entrada los valores de las variables de entrada en un tiempo *t-1*, como se muestra en

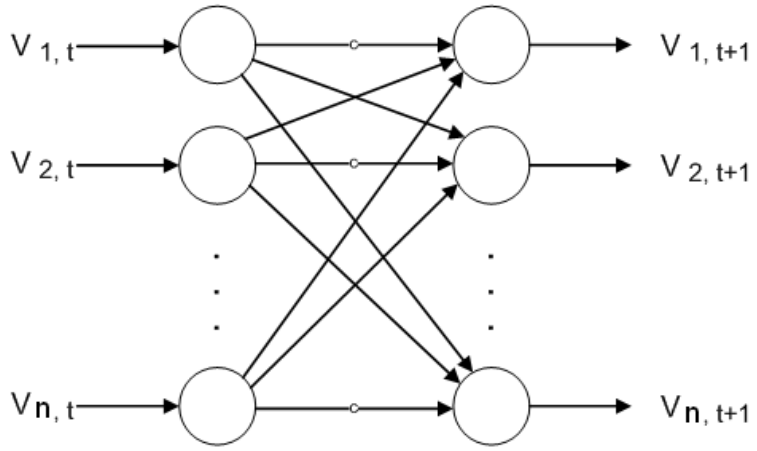

Figura 4.6.: Modelo de pronóstico basado en una red neuronal perceptron con n entradas y n salidas.

la tabla 4.2, y como salida su pronóstico en el tiempo t, como se muestra en la tabla 4.3. Los valores de un target i en el modelo de pronóstico, son los mismos valores de la variable de entrada i pero desfasados una unidad de tiempo. De tal forma que después del entrenamiento cuando la red tenga como entrada los valores de las variables en el tiempo t, pueda pronosticar su valor en el tiempo t+1.

Tabla 4.2.: Formato de los datos de entrada para el entrenamiento del modelo de pronostico.

tiempo	v_1	v_2	v_3	.	v_{n-2}	v_{n-1}	v_n
.	.	.	.	.	.	.	.
t-4	$v_{1,t-4}$	$v_{2,t-4}$	$v_{3,t-4}$	.	$v_{n-2,t-4}$	$v_{n-1,t-4}$	$v_{n,t-4}$
t-3	$v_{1,t-3}$	$v_{2,t-3}$	$v_{3,t-3}$	.	$v_{n-2,t-3}$	$v_{n-1,t-3}$	$v_{n,t-3}$
t-2	$v_{1,t-2}$	$v_{2,t-2}$	$v_{3,t-2}$	.	$v_{n-2,t-2}$	$v_{n-1,t-2}$	$v_{n,t-2}$
t-1	$v_{1,t-1}$	$v_{2,t-1}$	$v_{3,t-1}$	.	$v_{n-2,t-1}$	$v_{n-1,t-1}$	$v_{n,t-1}$
t	$v_{1,t}$	$v_{2,t}$	$v_{3,t}$	.	$v_{n-2,t}$	$v_{n-1,t}$	$v_{n,t}$

Este modelo VAR basado en una RN tipo Perceptron además de permitirnos pronosticar el valor de un conjunto de variables para utilizarlas como entrada de un modelo de predicción, puede apoyar al análisis del comportamiento fututo de variables de una

Tabla 4.3.: Formato de los datos de los targets para el entrenamiento del modelo de pronostico.

tiempo	$target_1$	$target_2$	$target_3$		$target_{n-2}$	$target_{n-1}$	$target_n$
.	.	.	.	.	.	.	.
t-4	$d_{1,t-3}$	$d_{2,t-3}$	$d_{3,t-3}$	.	$d_{n-2,t-3}$	$d_{n-1,t-3}$	$d_{n,t-3}$
t-3	$d_{1,t-2}$	$d_{2,t-2}$	$d_{3,t-2}$	.	$d_{n-2,t-2}$	$d_{n-1,t-2}$	$d_{n,t-2}$
t-2	$d_{1,t-1}$	$d_{2,t-1}$	$d_{3,t-1}$	.	$d_{n-2,t-1}$	$d_{n-1,t-1}$	$d_{n,t-1}$
t-1	$d_{1,t}$	$d_{2,t}$	$d_{3,t}$	.	$d_{n-2,t}$	$d_{n-1,t}$	$d_{n,t}$
t	$d_{1,t+1}$	$d_{2,t+1}$	$d_{3,t+1}$	.	$d_{n-2,t+1}$	$d_{n-1,t+1}$	$d_{n,t+1}$

red de telecomunicaciones.

4.2.6. Predicción del tráfico de la red.

Los valores de las variables que se pronostican con el modelo VAR son entradas del modelo de predicción. Este modelo predice el tráfico de la red basándose en las variables pronosticadas.

El modelo de predicción se desarrollo con una arquitectura de red neuronal. La figura 4.7 muestra la arquitectura general de una red MLP con n unidades de entrada (variables independientes), una capa oculta y una solo variable a predecir (variable dependiente).

La forma de entrenar una red neuronal MLP consiste en tener como datos de entrada los valores de las variables de entrada en el tiempo t+1, y como salida la predicción del target en el tiempo t+1.

En este paso de la simulación es posible tener una predicción del tráfico en un enlace de la red de datos. Además, ahora se conocen las variables que están relacionadas con la variable a predecir, de tal manera que en un escenario real se podría ir a coleccionar las estadísticas de estas variables para predecir de manera real el tráfico de un enlace dentro de la red de datos.

4.2.7. Interpretación de los resultados.

Una red se entrena minimizando una función de error (también llamado criterio de estimación). Varias funciones de error se basan en el principio de máxima probabilidad.

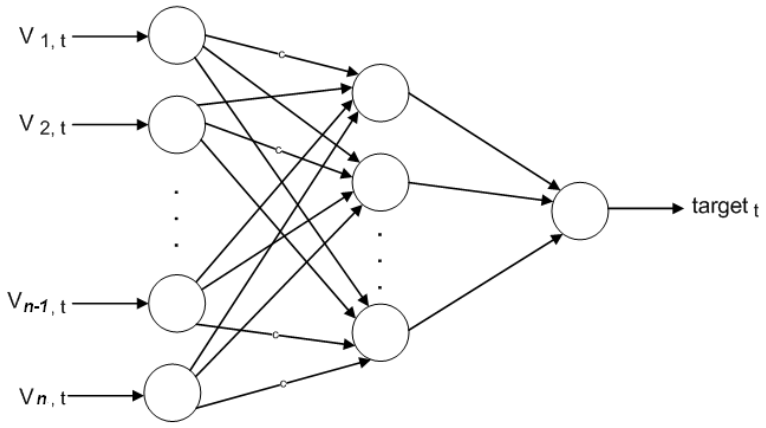

Figura 4.7.: Modelo de predicción basado en Redes Neuronales MLP.

La probabilidad se basa en una familia de distribuciones de error (ruido). Para variables de intervalo se ha utilizado la función de error normal.

La distribución normal también llamado criterio de error cuadrático medio. Frecuentemente utilizada para targets intervalo que tienen una distribución de ruido normal con varianza constante.

El propósito es encontrar una red que tenga buen desempeño con un conjunto de datos nuevos, la comparación más simple entre diferentes RN es evaluar la función de error usando datos que son independientes de los utilizados para el entrenamiento. Varias RN se entrenan minimizando una función de error definida con respecto al conjunto de datos de entrenamiento. El desempeño de las redes se compara evaluando la función de error sobre un conjunto independiente de datos de validación, y se selecciona a la red que tenga el error más pequeño con respecto a los datos de validación.

Para conocer el pronóstico del modelo es necesario calcular el error entre el valor que se pronostica y el valor real. Las siguientes métricas son utilizadas:

Suma de los errores cuadráticos

$$SSE = \sum_{t=1}^{n} (y_t - y'_t)^2 \tag{4.1}$$

Error cuadrático medio

$$MSE = \frac{SSE}{n} \tag{4.2}$$

La raíz del error cuadrático medio

$$RMSE = \sqrt{MSE} \tag{4.3}$$

Donde:

y_t es la salida (valor del pronóstico),
y'_t es el valor real,
n es el número de muestras.

4.3. Discusión

Después de aplicar la metodología propuesta para la predicción del tráfico en una red de datos, se observó que durante cada paso se van generando resultados parciales que nos permiten ir conociendo el comportamiento de la red de datos. Después de eliminar ruido en los datos es posible hacer análisis estadístico sobre las variables de la red para conocer su comportamiento. Sin embargo, si se quiere conocer la relación que tienen las variables para predecir el tráfico en la red, es necesario hacer un análisis de correlación y selección para identificar aquellas variables no redundantes que contribuyen mejor con la predicción del tráfico.

La ventaja más importante hasta este paso de la metodología es que se delimito el número de variables que se tienen que coleccionar en un escenario de red real para realizar el pronóstico de variables. Una contribución fue utilizar una red perceptron como modelo de pronóstico, que implementa un modelo VAR, el cual generalmente se aplica en el pronóstico de series de tiempo de variables económicas. Por último se estableció un modelo de predicción basado en RN, donde las entradas son las variables que se pronosticaron con el modelo VAR.

Capítulo 5.

Caso de Estudio

Se propone un escenario de red para demostrar la aplicación de la metodología propuesta. La metodología propuesta en este trabajo de tesis consiste en un conjunto de pasos aplicados para conocer la predicción de tráfico de los enlaces en una red de datos. En este capítulo se van aplicar cada uno de estos pasos a un caso de estudio, que consiste en un escenario de red de datos que es muy común en algunas empresas [4].

5.1. Coleccionar los datos de un escenario de red

Los datos que se coleccionan son las estadísticas generadas en los dispositivos de una red de datos. Esta red de datos es un escenario que se desarrolla en el simulador Opnet Modeler. Este escenario se basó en una topología de red que es frecuentemente utilizada por varias empresas y consiste en tener distribuidas varias subredes en un área geográfica, en donde todos los servicios de red están concentrados en un punto de la red [4].

5.1.1. Diseño de la red.

Este escenario consiste en una red de datos de una empresa en México que tiene 4 subredes distribuidas en los estados de Monterrey, Sinaloa, Chihuahua y Durango.

Con una topología en estrella las redes de Sinaloa, Chihuahua y Durango, se conectan a la red de Monterrey, la idea consiste en que todas las sucursales que se encuentren en el norte del país se conecten directamente a la red de Monterrey. La red de Monterrey se conecta de manera directa con la red de la Ciudad de México. En este escenario de red la información y servicios que proporciona la empresa se solicitan a la subred de la Ciudad de México.

La topología de red descrita anteriormente se muestra en la figura 5.1. Este escenario de red tiene cuatro enlaces de datos:

1. Entre las subredes de Sinaloa y Monterrey
2. Entre las subredes de Chihuahua y Monterrey
3. Entre las subredes de Durango y Monterrey
4. Entre las subredes de Monterrey y la Ciudad de México

Estos cuatro enlaces son DS1 PPP (1.544 Mbps). Cada una de las cuatro subredes esta compuesta por varias redes LAN. Todas interconectadas por un switch Cisco 2948G. El switch se conecta a un router Cisco 7204, que será el punto de conexión con cualquier subred. La figura 5.2 muestra los componentes de la subred de Sinaloa.

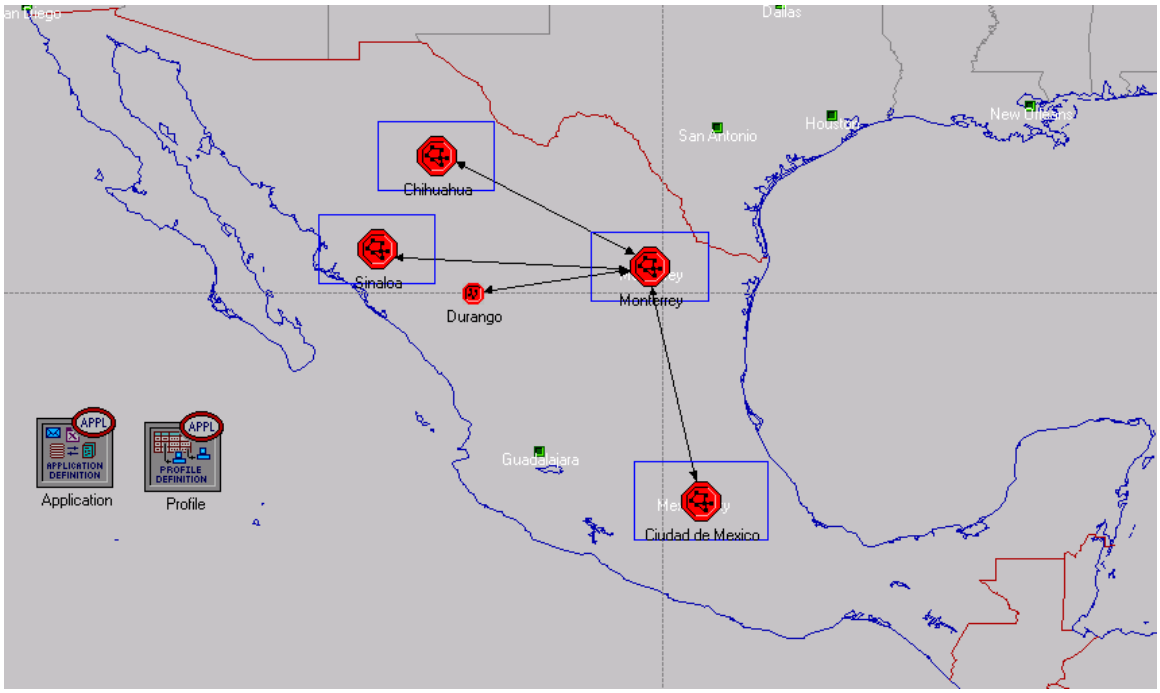

Figura 5.1.: Escenario de red.

Los dispositivos de la subred de la Ciudad de México son los servidores de la red, como se muestra en la figura 5.3. Estos servidores son el servidor de base de datos,

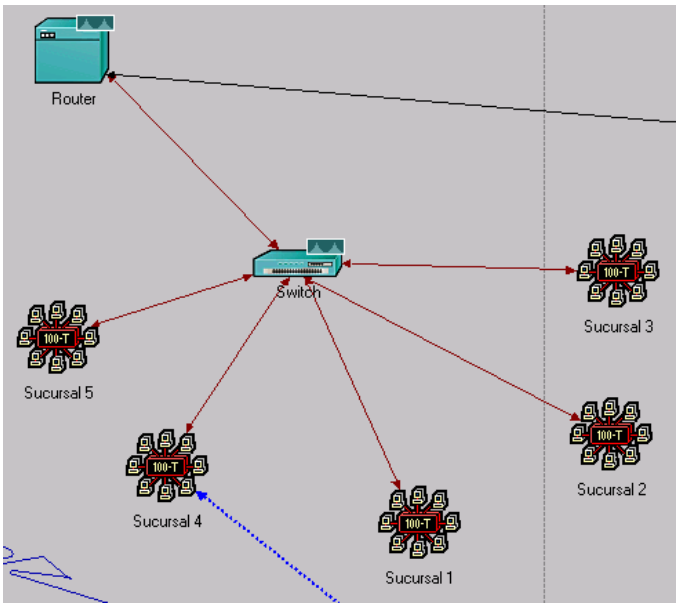

Figura 5.2.: Dispositivos de la subred de Sinaloa.

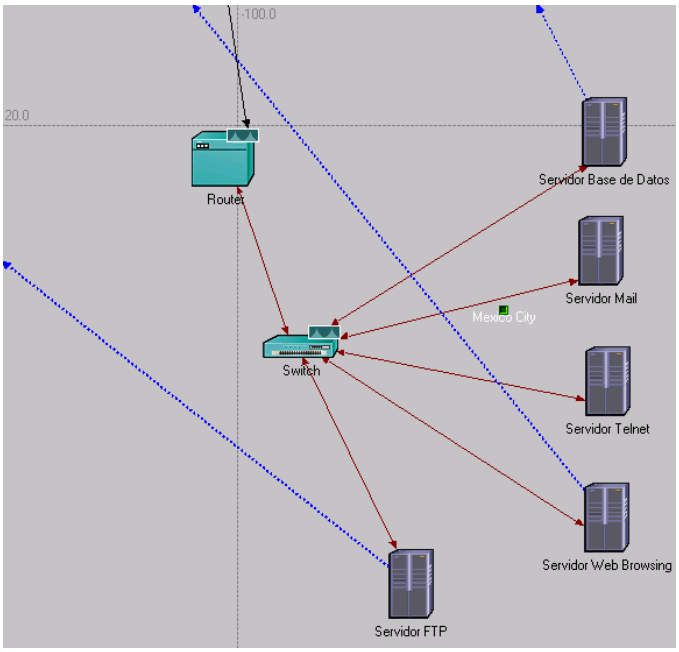

Figura 5.3.: Servidores de la subred de la Ciudad de México.

servidor ftp, servidor de correo, servidor telnet y el servidor de red.

Ahora que se ha creado la red se necesita especificar el tipo de tráfico. Una forma de simular tráfico en la red es utilizar flujo de tráfico IP (IP Traffic Flow) que consiste en especificar un nivel de tráfico entre dos subredes especificas. Por ejemplo si queremos una carga de tráfico de 500 kbps entre el servidor de base de datos y una subred de la ciudad de Durango. El nivel del flujo de tráfico IP se puede configurar para variar durante la simulación.

Se utilizo flujo de tráfico IP en los enlaces que se muestran en la tabla 5.1.

Tabla 5.1.: Tráfico de red.

Flujo de tráfico IP	**Enlace**
250,000 bps	Entre el servidor de base de datos y una red LAN de Durango.
250,000 bps	Entre el servidor web y una red LAN de Chihuahua.
250,000 bps	Entre el servidor FTP y una red LAN de Sinaloa.

El tráfico que se simula en el escenario de red depende del tipo de aplicaciones que se utilicen en la red. La tabla 5.2, muestra las aplicaciones y número de usuarios por cada una de las subredes. Por ejemplo 6 usuarios de la subred de Sinaloa pueden consultar el servidor de base de datos de la Ciudad de México.

La carga de tráfico dependerá en gran medida de las consultas a la base datos, a las transacciones a través de FTP y a la consulta de imágenes. Las consultas a la base de datos serán de 32 kbytes con una distribución exponencial entre los tiempos de llegada de las consultas y con una atención del 100 % de las consultas. La archivos enviados por medio de FTP tienen tamaño de 50 kbytes y el tiempo de transferencia entre archivos tiene una distribución exponencial. La configuración de las aplicaciones las establece Opnet Modeler [3].

5.1.2. Generación de datos de tráfico

Antes de iniciar la simulación es necesario especificar las estadísticas que se quieren coleccionar. Estas estadísticas permitirán conocer el desempeño de la red. Las estadísticas de red seleccionadas de acuerdo a la clasificación de Opnet Modeler y a los propósitos de este trabajo de investigación son:

Tabla 5.2.: Aplicaciones y perfiles de las subredes.

Subred	**Aplicación**	**Número de usuarios en la subred**
Sinaloa	Http (Image browsing)	6
Sinaloa	Http (Web browsing)	150
Sinaloa	FTP	150
Sinaloa	DB Query	6
Durango	Http (Image browsing)	6
Durango	Http (Web browsing)	150
Durango	FTP	150
Durango	DB Query	4
Chihuahua	Http (Image browsing)	3
Chihuahua	Http (Web browsing)	150
Chihuahua	FTP	150
Chihuahua	DB Query	8
Monterrey	Http (Image browsing)	6
Monterrey	Http (Web browsing)	150
Monterrey	FTP	150
Monterrey	DB Query	4

- Estadísticas globales: Data Base, Email, Ftp, Ethernet and http.
- Estadísticas por nodo: Client DB, Client Email, Client Ftp, Client http, CPU, Ethernet Channel, Ethernet, LAN, Server DB, Server Email, Server Ftp, Server http, Server performance and Switch.
- Estadísticas por enlace: Queuing delay, throughput and utilization.

Las estadísticas globales permiten conocer el efecto que tienen algunas aplicaciones en toda la red de datos. Sin embargo, nuestro propósito es seleccionar variables que sean factibles de coleccionar en una red de datos real. De tal forma, que para el análisis de predicción solo se consideraron estadísticas de los nodos y enlaces de la red.

Hay dos propiedades importantes que se tienen que definir en la simulación: la *duración* y el *número de valores por estadística*. Se estableció una duración de 500 minutos y 5000 valores por estadística, de tal manera que el tiempo entre cada valor de la estadística está definido por la siguiente ecuación:

$$T = \frac{D}{M} \tag{5.1}$$

Donde:
T es el tiempo entre cada valor de la estadística en segundos,
D es la duración de la simulación en segundos,
M es el número de muestras por estadística,

Para nuestra simulación sería:

$$T = \frac{30000}{5000} = 6 \tag{5.2}$$

Las valores en las estadísticas estarán igualmente espaciadas a 6 segundos.

El objetivo es la predicción de tráfico en redes de datos, para ello es necesario analizar las estadísticas que contribuyan con los cambios en los niveles de utilización de los enlaces de datos. El análisis se concentro en el enlace entre las subredes de la Ciudad de México y Monterrey. La figura 5.4 muestra su comportamiento durante el tiempo de simulación y proporciona la siguiente información:

- Un mínimo de 40 % y un máximo de 100 % en la utilización del enlace.
- El porcentaje de utilización del enlace no tiene un comportamiento periódico.

De acuerdo al número de dispositivos utilizados en el escenario de red y a las estadísticas seleccionadas se obtuvieron 191 estadísticas. Opnet Modeler permite exportar las estadísticas generadas en la simulación. Cada estadística es una variable de red. Con las estadísticas de red se puede formar una serie de tiempo con 191 variables.

Parte de esta serie de tiempo se muestra en la tabla 5.3, la primera columna representa la variable de tiempo, que de acuerdo a sus valores muestra una serie de tiempo igualmente espaciada en el tiempo, con intervalos de 6 segundos. Las columnas v_1, v_2,..., v_{191} representan las 191 variables de red.

Como se muestra en la tabla 5.3, algunas de las estadísticas de la simulación no tienen valores definidos, por tal motivo no es posible conocer que tipo de variables se tienen (unarias, binarias, nominales, ordinales). Este problema se soluciona con la limpieza de los datos.

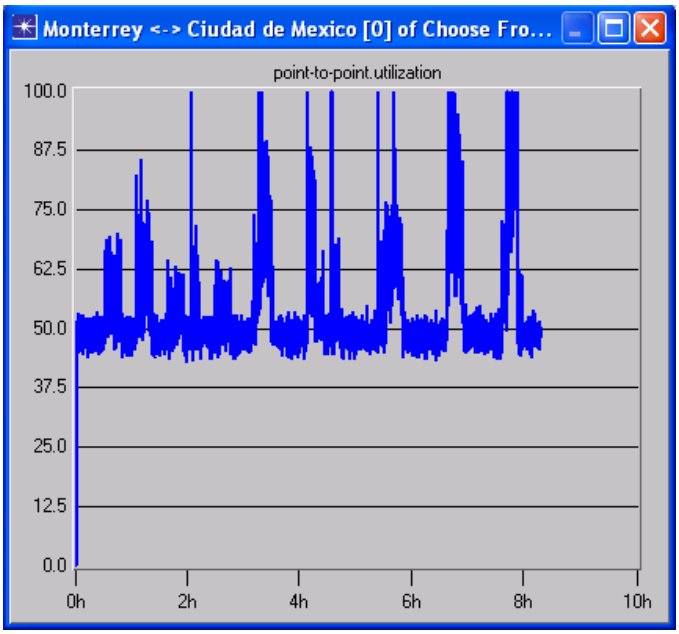

Figura 5.4.: Estadística de la utilización del enlace de datos entre las subredes de Monterrey y la Ciudad de México.

Tabla 5.3.: Serie de tiempo.

Tiempo (seg)	v_1	v_2	.	v_{190}	v_{191}
.	.	.	.	.	.
1956	#N/A	0.74028	.	#N/A	#N/A
1962	#N/A	0.77433	.	#N/A	#N/A
1968	#N/A	0.76733	.	#N/A	#N/A
1974	#N/A	0.77653	.	#N/A	#N/A
1980	0.00012	0.76483	.	#N/A	0.000167
1986	#N/A	0.75782	.	#N/A	#N/A
1992	0.01087	1.00927	.	0.000114	3.31678
1998	0.00167	0.77945	.	0.000134	0.55674
2004	0.00360	0.83691	.	0.000194	1.10432
2010	0.00418	0.72172	.	0.000115	0.00318
.	.	.	.	.	.

5.2. Preparación y limpieza de los datos.

El procesamiento de los datos se realiza sobre la serie de tiempo construida. El primer paso consiste en analizar ruido e inconsistencia en los datos. Se elimino el problema de valores no definidos utilizando el programa del apéndice B. La tabla 5.4 muestra una parte de la serie de tiempo después del proceso de limpieza.

Tabla 5.4.: Serie de tiempo después de la limpieza de los datos.

Tiempo (seg)	v_1	v_2	.	v_{190}	v_{191}
.	.	.	.	.	.
1956	0	0.74028	.	0	0
1962	0	0.77433	.	0	0
1968	0	0.76733	.	0	0
1974	0	0.77653	.	0	0
1980	0.00012	0.76483	.	0	0.000167
1986	0	0.75782	.	0	0
1992	0.01087	1.00927	.	0.000114	3.31678
1998	0.00167	0.77945	.	0.000134	0.55674
2004	0.00360	0.83691	.	0.000194	1.10432
2010	0.00418	0.72172	.	0.000115	0.00318
.	.	.	.	.	.

Con la eliminación de ruido en los datos, se puede obtener la siguiente información:

- Se detectaron el siguiente tipo de variables:

  - Variables clase (75): Incluye variables unarias, binarias y ordinales. Esto se debe a que se seleccionaron estadísticas que durante la simulación no tuvieron valores. Por ejemplo: la carga en el servidor de base de datos (solicitudes/sec) de consultas de 512 bytes es cero, ya que ningún usuario durante la simulación hizo este tipo de consultas.

  - Variables intervalo (116): Son variables que tuvieron mas de 10 valores numéricos. Ejemplo: el Tráfico enviado del canal ethernet del servidor FTP (bits/sec).

- Información de la variable que mide el porcentaje de utilización del enlace de la Ciudad de México y Monterrey: valor mínimo 0 %, valor máximo 100 %, media 53.574 % y desviación estándar 11.682 %.
- Representación gráfica de la distribución de los datos de la variable utilización, ver figura 5.5, donde se muestra que la mayor parte de los valores de utilización del enlace se encuentran cerca del 50 %.

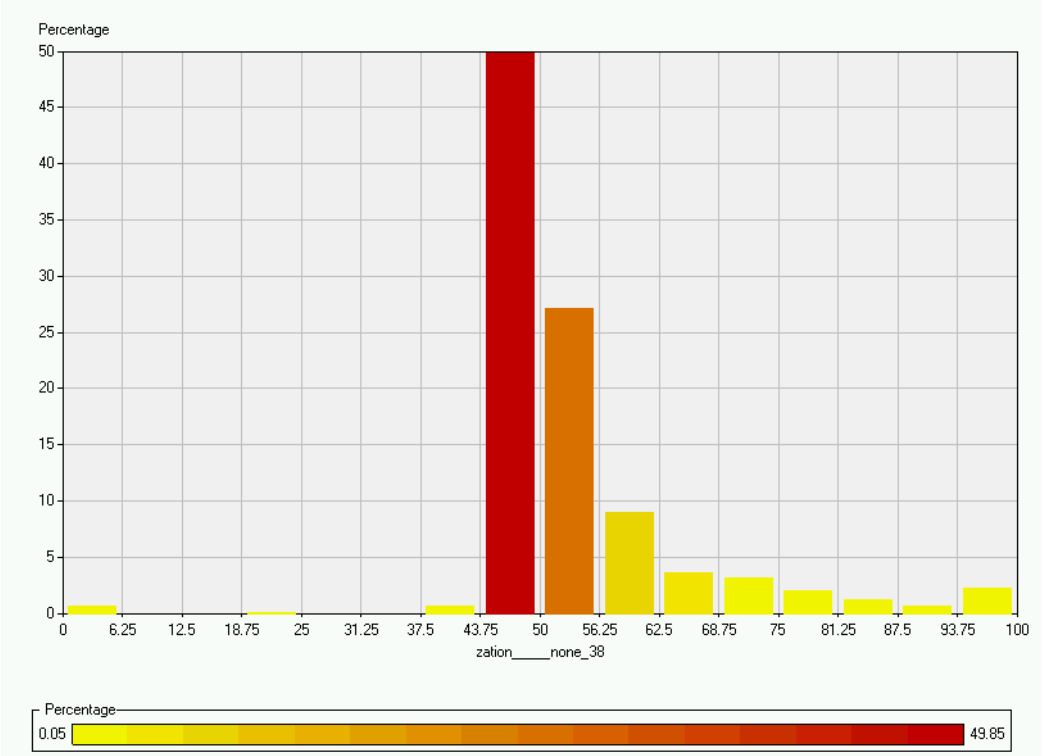

Figura 5.5.: Distribución del porcentaje de utilización del enlace de la Ciudad de México y Monterrey.

Para llevar acabo el análisis estadístico de las variables de la serie de tiempo, se utilizó una herramienta llamada SAS Enterprise Miner, que consiste de un número de nodos para limpiar los datos, explorar y analizar datos, desarrollo y validación de modelos de Minería de Datos [39].

Sin embargo un análisis estadístico solo permite obtener información de la utilización del enlace, aun no se conoce la relación o dependencia que tiene con las demás variables.

5.3. Selección de variables no redundantes

Después de tener consistencia en los datos, se inicio el análisis de variables en la serie de tiempo. El primer paso consiste en detectar un grupo de variables no redundantes, con la finalidad de disminuir el número de variables pero manteniendo la misma consistencia en los datos. Este problema se resolvió utilizando la correlación de Spearman.

Para llevar acabo la correlación de Spearman entre las 191 variables de la serie de tiempo, se utilizo SAS Enterprise Miner. Como resultado obtuvimos una matriz de 191x191, con los valores de la correlación entre las 191 variables, como se muestra en la tabla 5.5.

En la simulación se selecciono la estadística de utilización de todos los enlaces de la red de datos. Sin embargo después de hacer el análisis de correlación se detecto un problema, y consistía que las variables no redundantes de la red eran todas las estadísticas de utilización de todos los enlaces. De acuerdo a los objetivos de la tesis, no sería viable que, para predecir el nivel de utilización de un enlace se tuviera que predecir la utilización de algún otro. Por tal motivo en el análisis de correlación se descartaron todas la variables de utilización de los enlaces, y solo se considero la variable utilización del enlace que se quiere predecir, en este caso el enlace entre las subredes de la Ciudad de México y Monterrey.

Tabla 5.5.: Parte de la matriz de correlación de Spearman entre las 191 variables de la serie de tiempo.

	v_1	v_2	.	v_{169}	v_{170}	.	v_{190}	v_{191}
v_1	1	0.02782	.	0.03996	0.04	.	-0.0977	-0.07305
v_2	.	1	.	0.80259	0.80259	.	0.65249	0.63047
.	.	.	.	.	.	.	.	.
v_{169}	.	.	.	1	**1**	.	0.5029	0.45999
v_{170}	.	.	.	.	1	.	0.50298	0.4601
.	.	.	.	.	.	.	.	.
v_{190}	.	.	.	.	.	.	1	0.94218
v_{191}	.	.	.	.	.	.	.	1

Un ejemplo de variables redundantes es entre las variables v_{169} y v_{170}, esto se debe a que la v_{169} es el porcentaje de utilización del enlace entre las subredes de la Ciudad de

México y Monterrey y la v_{170} es el throughput del enlace entre las mismas subredes.

El resultado de la correlación entre estas dos variables es de 1, por tal motivo para tener la información del enlace solo basta con considerar alguna de las dos variables.

Se estableció un umbral de correlación u de 0.8. De esta manera se analizo la matriz de correlación y aquellas variables que tuvieran una correlación mayor al umbral u, se consideraban como redundantes y no serían seleccionadas, este análisis se realizó con el programa del apéndice B. La tabla 5.6 muestra las 19 variables no redundantes seleccionadas por el análisis de correlación.

Tabla 5.6.: Variables no redundantes.

1	Carga de sesiones por segundo en el servidor FTP.
2	Porcentaje de utilización del CPU del servidor de base de datos (%).
3	Tráfico enviado del canal ethernet del servidor de base de datos (bits/sec).
4	Carga en el servidor de base de datos (sesiones/sec).
5	Carga en el servidor de base de datos (tareas/sec).
6	Cantidad de retransmisiones TCP del servidor de base de datos.
7	Porcentaje de utilización del CPU del servidor FTP (%).
8	Tráfico enviado del canal ethernet del servidor FTP (bits/sec).
9	Tráfico recibido del canal ethernet del servidor FTP (bits/sec).
10	Carga en el servidor FTP (sesiones/sec).
11	Cantidad de retransmisiones TCP del servidor FTP.
12	Porcentaje de utilización del CPU del servidor de correo (%).
13	Tráfico recibido del canal ethernet del servidor de correo (bits/sec).
14	Porcentaje de utilización del CPU del servidor web (%).
15	Tráfico enviado del canal ethernet del servidor web (bits/sec).
16	Tráfico recibido del canal ethernet del servidor web (bits/sec).
17	Carga en el servidor web (Imágenes)(solicitudes/sec).
18	Cantidad de retransmisiones TCP del servidor web.
19	Porcentaje de utilización del enlace entre las subredes de la Ciudad de México y Monterrey (%).

La figura 5.6, muestra la gráfica del número de variables no redundantes obtenidas variando el nivel de umbral. Se puede ver que en el escenario de red del caso de estudio

existe una gran cantidad de variables redundantes.

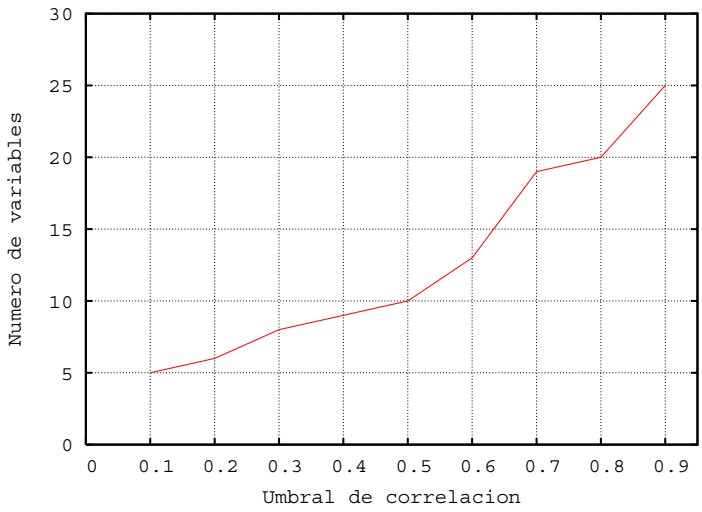

Figura 5.6.: Gráfica de correlación del número de variables no redundantes variando el valor del umbral de 0.1 a 0.9.

La tabla 5.6 demuestra que el tráfico generado en la red de telecomunicaciones depende en gran medida de los servicios proporcionados por los servidores de la Ciudad de México. Podemos agrupar las variables seleccionadas en 3 tipos: Carga de sesiones por segundo en los servidores, porcentaje de utilización del CPU de los servidores, tráfico enviado del canal ethernet de los servidores (bits/sec) y el tráfico recibido del canal ethernet de los servidores (bits/sec).

Si el análisis de las variables de red hubiera iniciado sin conocer con anterioridad la topología de la red ni el tipo de servicios que se ofrecen, con los resultados obtenidos hasta el momento se podría saber que los servicios proporcionados por la red están concentrados en un solo punto de la red y que los servicios que proporciona la red son aplicaciones FTP, consultas a base de datos, consultas a servidores de red y correo. Además que en gran medida estas aplicaciones son las que generan el tráfico de la red.

En esta etapa de la metodología se demuestra que un proceso de adquisición de conocimiento nos permite obtener información de los datos sin tener conocimiento pre-

vio de ellos. Con esto se demuestra que un análisis de correlación puede ayudar a seleccionar variables importantes en la red, pero aun sin conocer la relación o dependencia que hay entre ellas. Ahora la siguiente etapa de la metodología consiste en seleccionar de las 19 variables no redundantes cuales de ellas están relacionadas o contribuyen con los cambios en el porcentaje de utilización del enlace entre las subredes de la Ciudad de México y Monterrey.

5.4. Selección de variables para la predicción

Para seleccionar las variables que estén relacionadas con la variable de tráfico se desarrolla un modelo de selección basado en árboles de decisión. SAS Enterprise Miner proporciona un nodo de árboles de decisión para resolver este problema, como se muestra en la figura 5.7. Este modelo esta compuesto por dos nodos:

1. El primer nodo es el conjunto de datos, integrado por las 19 variables seleccionadas anteriormente. SAS Enterprise Miner permite importar y exportar archivos de datos.

2. El segundo es un nodo de árbol de decisión. Este nodo forma el árbol de decisión de acuerdo al criterio de división seleccionado.

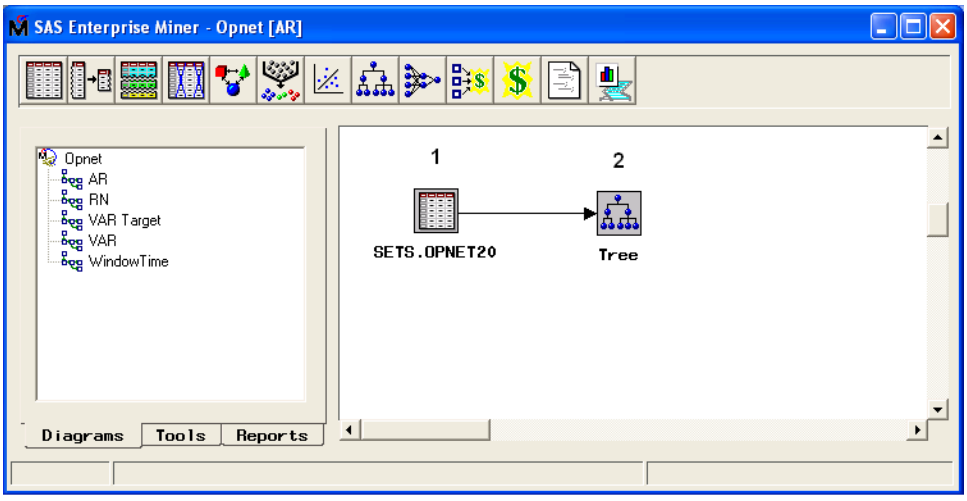

Figura 5.7.: Diagrama del modelo de selección basado en árboles de decisión.

El modelo selecciono 7 variables, que se listan en la tabla 5.7. La primera columna consiste en el nivel de importancia que tiene cada variable con respecto al target. Se

muestra que la variable con mayor relación con el target es el tráfico recibido del canal ethernet del servidor web (bits/sec), lo que indica que el tráfico en el enlace de red de la Ciudad de México y Monterrey depende en gran medida del tráfico generado por las consultas al servidor web.

En general de acuerdo a la información que nos proporciona la tabla 5.7, podemos ver que hay al menos una variable seleccionada por cada uno de los servicios que proporciona la red.

Tabla 5.7.: Variables que contribuyen con el target.

	Importancia	**Variable**
	Target	Porcentaje de utilización del enlace de datos entre la Ciudad de México y Monterrey (%).
v_1	**1.0**	Tráfico recibido del canal ethernet del servidor web (bits/sec).
v_2	**0.6618**	Carga en el servidor FTP (sesiones/sec).
v_3	**0.5319**	Tráfico enviado del canal ethernet del servidor FTP (bits/sec).
v_4	**0.3851**	Porcentaje de utilización del CPU del servidor de base de datos (%).
v_5	**0.1870**	Porcentaje de utilización del CPU del servidor de web (%).
v_6	**0.1836**	Tráfico enviado del canal ethernet del servidor web (bits/sec).
v_7	**0.1323**	Tráfico enviado del canal ethernet del servidor de base de datos (bits/sec).

La gráfica de la variable v_1 durante todo el tiempo de la simulación se muestra en la figura 5.8. Recordemos que el servicio web se le brinda a 450 usuarios, lo que muestra la gráfica es que durante algunos instantes de la simulación varios usuarios coinciden en la utilización del servicio web. Con estos resultados y debido a la importancia que tiene esta variable con respecto al target, se podría decir que generalmente cuando hay varios usuarios haciendo consultas al servidor web es cuando se genera parte del tráfico en el enlace entre las subredes de la Ciudad de México y Monterrey. Alrededor de los 28000 segundos, es cuando se genera la mayor cantidad de tráfico recibido en el servidor web.

La gráfica de la variable v_2 durante todo el tiempo de la simulación se muestra en la figura 5.9. Recordemos que el servicio de FTP se le brinda a 450 usuarios, lo que mues-

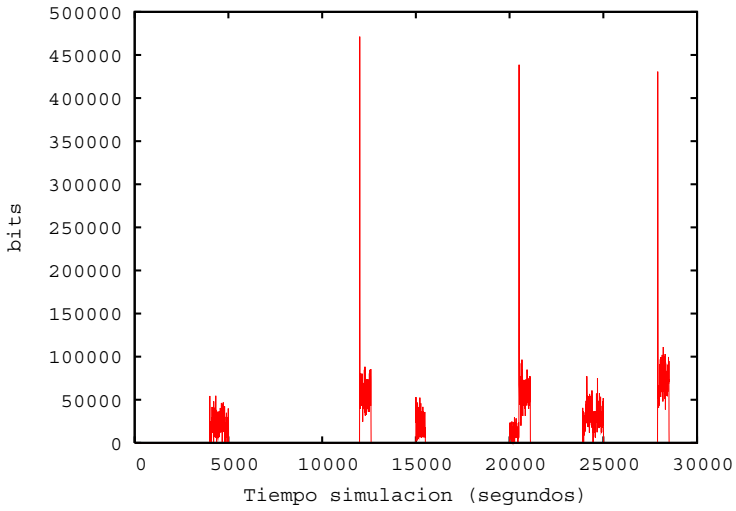

Figura 5.8.: Tráfico recibido del canal ethernet del servidor web (bits/sec).

tra la gráfica es que el tiempo que hacen uso de este servicio es durante periodos muy cortos, pero por el nivel de importancia que tiene con respecto al target, es probable que coincida que los momentos en que se genera tráfico en el enlace que estamos analizando, sea cuando se estén generando la mayor cantidad de sesiones en el servidor FTP.

La gráfica de la variable v_3 durante todo el tiempo de la simulación se muestra en la figura 5.10. Se observa en la gráfica que gran parte del tráfico enviado por el servidor FTP se debe a la configuración que se hizo en la simulación, que consistió en generar 250 kbps de tráfico IP entre el servidor FTP y una red LAN de la ciudad de Sinaloa. Alrededor de los 12000 y 28000 segundos son los periodos en los que el servidor FTP envía más tráfico.

La gráfica de la variable v_4 durante todo el tiempo de la simulación se muestra en la figura 5.11. Recordemos que el servicio de consultas de 32 kbps a la base de datos se permitió a 22 usuarios. Se puede observar en la gráfica que no existe una utilización periódica del servidor de base de datos pero si muy frecuente. Alrededor de los 13000 segundos fue el momento en el se presento la mayor utilización del servidor. La importancia que tiene esta variable con respecto al target se debe a la frecuencia con la que se hace uso del servidor de base da datos.

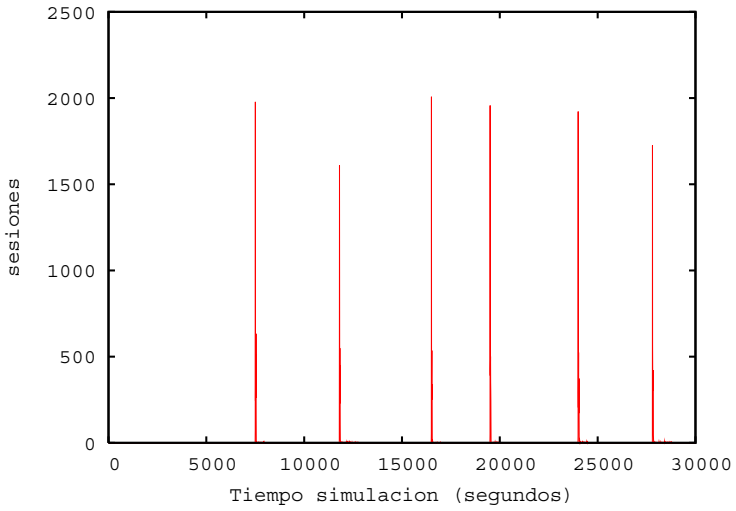

Figura 5.9.: Carga en el servidor FTP (sesiones/sec).

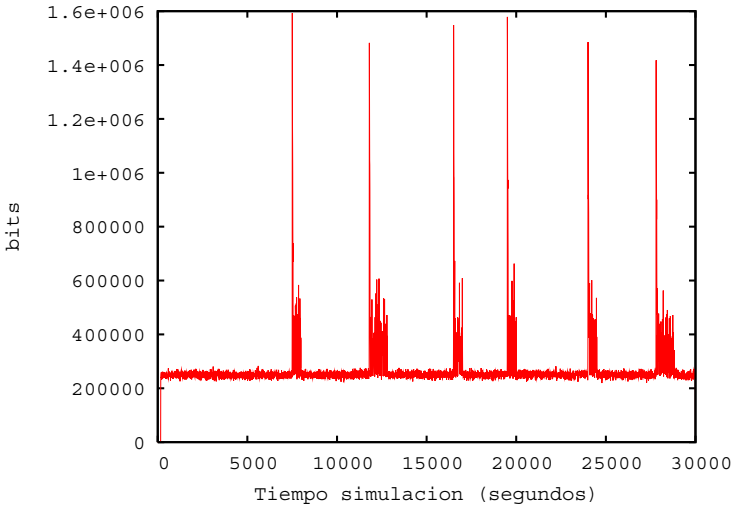

Figura 5.10.: Tráfico enviado del canal ethernet del servidor FTP (bits/sec).

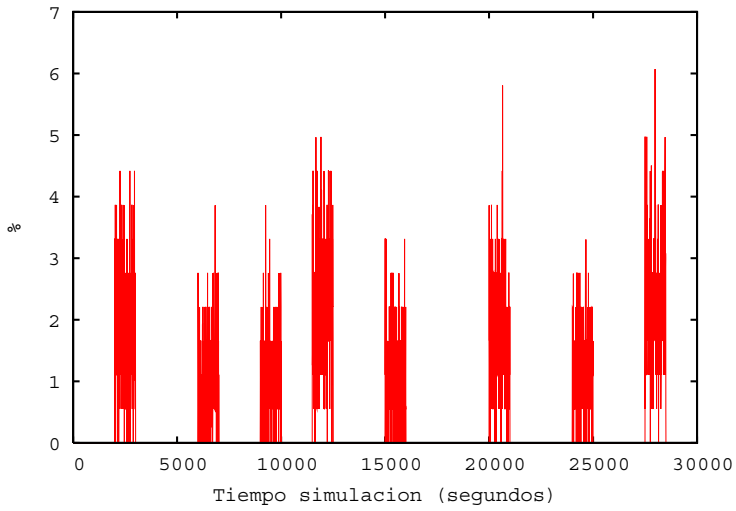

Figura 5.11.: Porcentaje de utilización del CPU del servidor de base de datos (%).

La gráfica de la variable v_5 durante todo el tiempo de la simulación se muestra en la figura 5.12. Se puede observar en la gráfica que no existe una utilización periódica del servidor web. Alrededor de los 24000 segundos fue el momento en el que se presento por mayor tiempo la utilización del servidor web. En el servidor web se presentaron niveles de utilización de hasta el 8 %. Sin embargo es mayor la frecuencia con la que se utiliza el servidor de base de datos.

La gráfica de la variable v_6 durante todo el tiempo de la simulación se muestra en la figura 5.13. Se puede observar en la gráfica que no existe un comportamiento periodico del tráfico enviado por el servidor web. Alrededor de los 28000 segundos fue el momento en el que se genero la mayor cantidad de tráfico por parte del servidor web. En el servidor web se presentaron niveles de tráfico enviado de hasta el 800 kbps comparado con los 400 kbps que presenta la gráfica del tráfico enviado por servidor de FTP 5.10. Con esta comparación se puede concluir que es mayor el tráfico generado por el servidor web.

La gráfica de la variable v_7 durante todo el tiempo de la simulación se muestra en la figura 5.14. Se puede observar en la gráfica que no existe un comportamiento periódico del tráfico enviado por el servidor de base da datos. En el servidor de base de datos

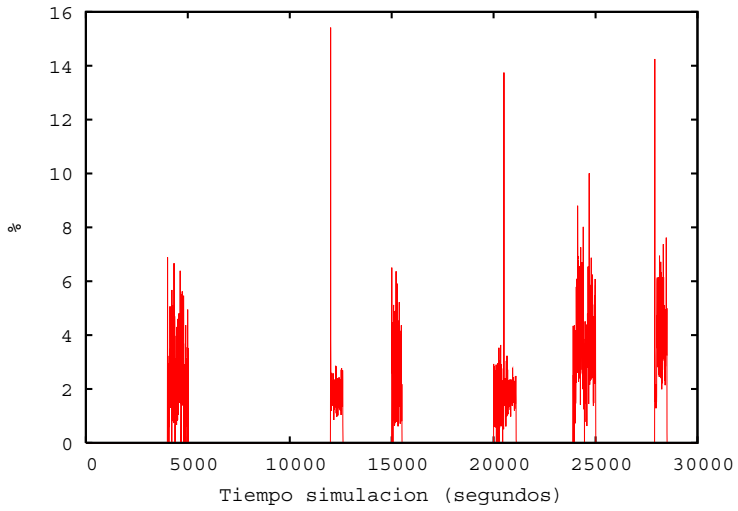

Figura 5.12.: Porcentaje de utilización del CPU del servidor de web (%).

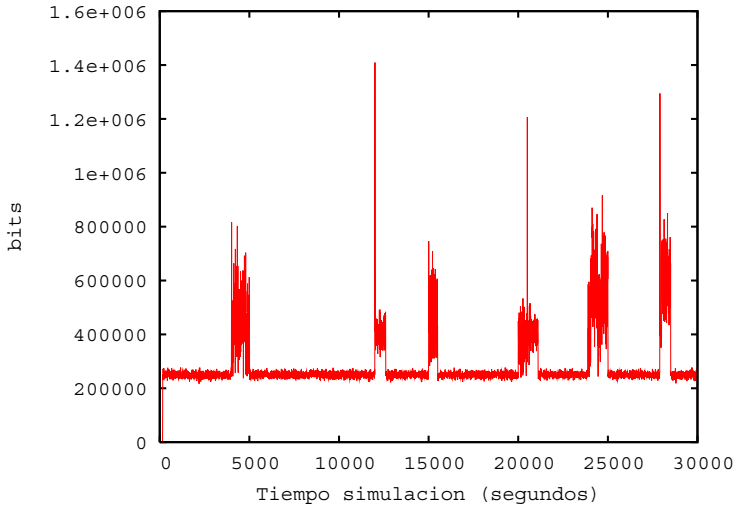

Figura 5.13.: Tráfico enviado del canal ethernet del servidor web (bits/sec).

se presentaron niveles de tráfico enviado de hasta el 600 kbps comparado con los 800 kbps que presenta la gráfica el servidor de web 5.13 y los 400 kbps del servidor de FTP 5.10. Con esta comparación se puede concluir que el servidor de aplicaciones que genera mayor tráfico en algún instante de tiempo es el servidor web, pero el servidor que genera tráfico con mayor frecuencia es el servidor de base de datos.

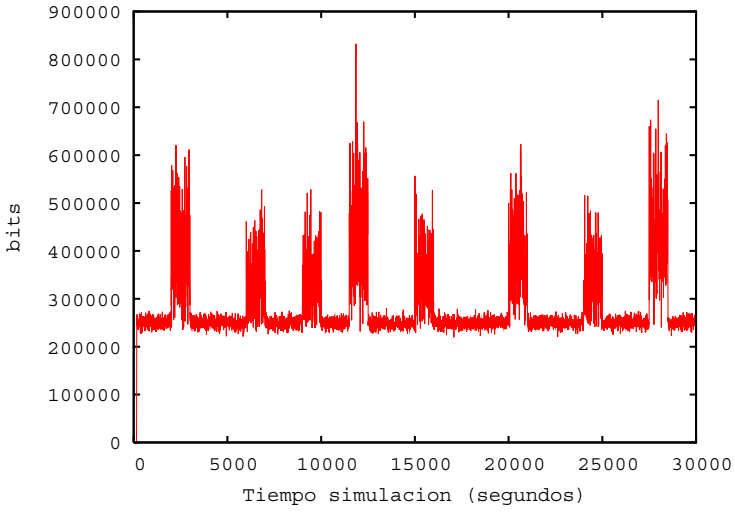

Figura 5.14.: Tráfico enviado del canal ethernet del servidor de base de datos (bits/sec).

Hasta esta etapa de la metodología se ha convertido los datos en información. El análisis inicio a partir de una serie de tiempo de 191 variables y ahora después del procesamiento se han seleccionado las 7 variables de la tabla 5.7. También podemos concluir hasta esta etapa que otra de las ventajas de la metodología es proporcionar información a partir de un conjunto de datos, sin importar si el análisis lo hace o no una persona experta en redes de datos.

Como resultado del análisis de las gráficas de las 7 variables seleccionadas se encontró que los instantes en los que se genera el tráfico de cada servicio no tienen un comportamiento periódico, además que cada variable tiene una distribución aleatoria del momento en que se genera el tráfico, de tal manera que es conveniente ver la relación que existe entre estas variables para la predicción del tráfico del enlace entre las subredes de la Ciudad de México y Monterrey.

5.5. Pronóstico de las variables seleccionadas.

Como mencionamos anteriormente solo tenemos datos hasta el tiempo t. Para predecir el target en el tiempo t+1 necesitamos los valores de las 7 variables de entrada en el tiempo t+1. Debido a que estos datos no existen en el tiempo de la predicción nosotros pronosticamos los valores de las 7 variables en el tiempo t+1 usando un modelo VAR, el cual se basa en sus propios valores retrasados y en los valores retrasados de las demás variables. Como se menciono en el capítulo 4, se puede desarrollar un modelo VAR por medio de una Red Neuronal Perceptron, ver figura 4.6. Donde las 7 variables seleccionadas anteriormente son entrada del modelo VAR en el tiempo t y la salida es el valor de las 7 variables en un tiempo t+1.

El modelo VAR se implementa en SAS Enterprise Miner por medio del nodo de Redes Neuronales (RN). El diagrama del modelo de pronostico se muestra en la figura 5.15.

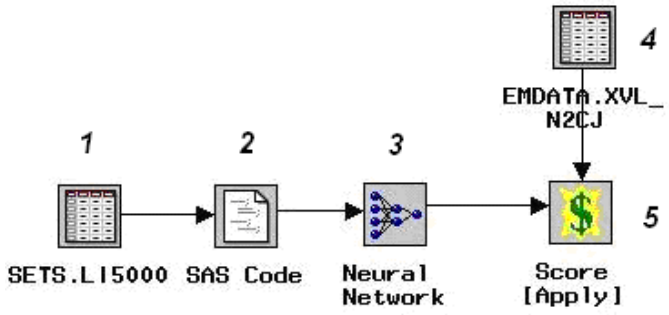

Figura 5.15.: Diagrama en SAS Enterprise Miner del modelo de pronostico.

Este modelo esta compuesto por cinco nodos:

1. El primer nodo es el conjunto de datos, integrado por las 7 variables seleccionadas anteriormente.

2. El 90 % de los datos se utilizaron para entrenar el modelo y el 10 % para su validación. Para llevar acabo el particionamiento de los datos SAS Miner cuenta con un nodo de particionamiento que particiona los datos tomando muestras de manera aleatoria dentro de todo el conjunto de datos, sin embargo el orden de las muestras de la serie de tiempo de las variables seleccionadas no se puede cambiar,

de tal manera que para conservar el orden cronológico de los datos se utiliza un nodo llamado SAS Code.

3. El tercer nodo es un nodo de RN. La estructura interna de este nodo consiste en una red neuronal Perceptron, como se muestra en la figura 5.16.

4. Datos de validación.

5. Evaluar el modelo con los datos de validación. Se obtienen los valores de las 7 variables de entrada en el tiempo t+1.

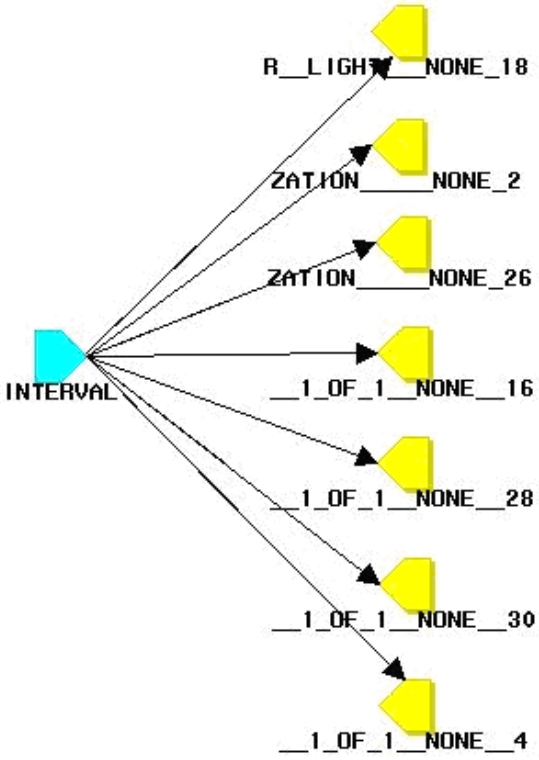

Figura 5.16.: Nodo de red neuronal en SAS Enterprise Miner que implementa una red neuronal Perceptron.

En esta etapa de la metodología se conoce las variables que están relacionadas con el target, y se ha pronosticado su valor en un tiempo t+1, para que puedan ser entrada de un modelo de predicción, donde se busca predecir el valor del target en un tiempo

t+1.

La tabla 5.8 presenta la SSM, SME, RSME del pronóstico de las 7 variables.

Tabla 5.8.: Medidas de desempeño del modelo de pronóstico.

Variable	**SSE**	**MSE**	**RMSE**
V_1	288076919216	576153838.43	24003.204753
V_2	3724233.8273	7448.4676546	86.304505413
V_3	3.432833E12	6865666028.5	82859.314675
V_4	392.87221341	0.7857444268	0.8864222621
V_5	448.15971174	0.8963194235	0.9467414766
V_6	2.6719506E12	5343901137.5	73101.991338
V_7	2.6153981E12	5230796158.6	72324.243228

A continuación se presentan las gráficas comparativas de los valores obtenidos de la simulación y del pronóstico. El pronóstico que se muestra en las graficas es sobre los datos de validación para cada una de las variables. Los valores de validación corresponden a los últimos 3000 segundos de la simulación

La gráfica 5.17 muestra una comparación de los valores obtenidos del modelo de pronóstico y los que se obtuvieron en la simulación de la variable V_1. La variable V_1, obtuvo un error (promedio) de pronóstico de 24.003 kbps entre los valores de la simulación y los obtenidos del pronóstico, como se muestra en la tabla 5.8. Se puede ver en la gráfica que se tiene dificultad en detectar los cambios abruptos en el tráfico.

La gráfica 5.18 muestra una comparación de los valores obtenidos del modelo de pronóstico y los que se obtuvieron en la simulación de la variable V_2. Se puede ver en la gráfica que durante el periodo de validación esta variable solo estuvo presente por un tiempo de alrededor de 100 segundos. La variable V_2, obtuvo un error (promedio) de pronóstico de 86.30 sesiones por segundo entre los valores de la simulación y los obtenidos del pronóstico, este bajo nivel de pronóstico que se muestra en la gráfica se debe a la poca cantidad de valores de esta variable en la etapa de entrenamiento.

La gráfica 5.19 muestra una comparación de los valores obtenidos del modelo de pronóstico y los que se obtuvieron en la simulación de la variable V_3. Se puede ver en la gráfica que durante el periodo de validación esta variable se presenta por un tiempo

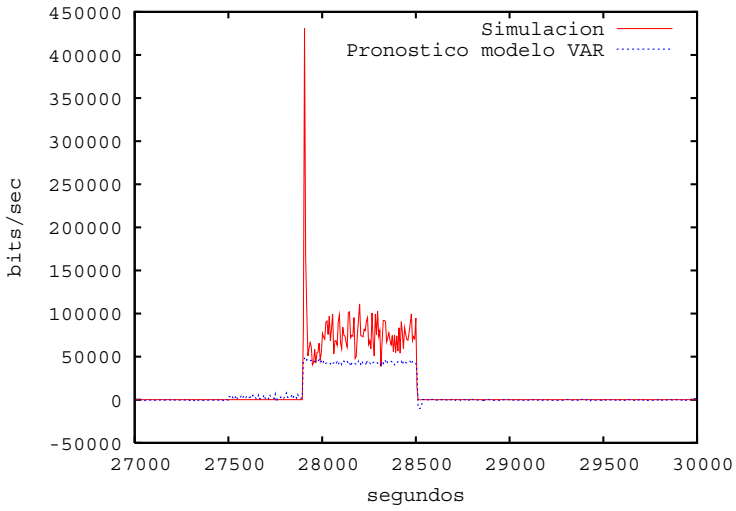

Figura 5.17.: Tráfico recibido del canal ethernet del servidor web (bits/sec).

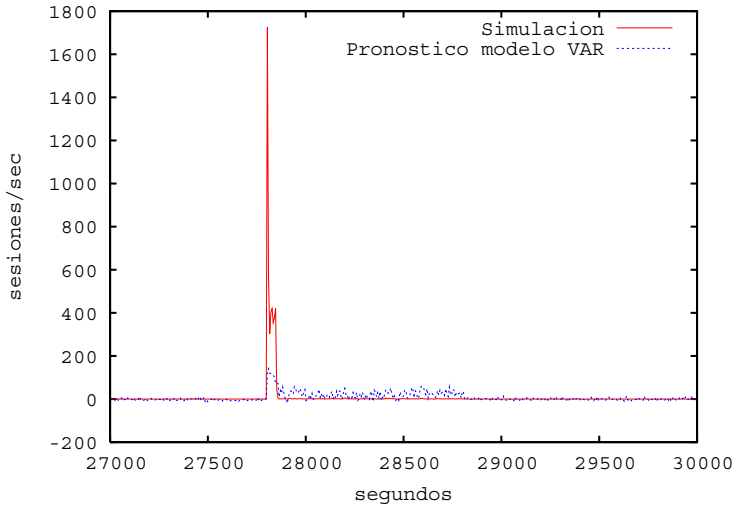

Figura 5.18.: Carga en el servidor FTP (sesiones/sec).

de alrededor de 1100 segundos. Comparado con la variable v_1, en esta variable si se logran pronosticar los cambios abruptos de tráfico. El error(promedio) de pronóstico de la variable V_3 es de 82.859 kbps. El error de pronóstico de esta variable es mayor al de la variable V_1, debido a que se tiene más tráfico en el servidor FTP.

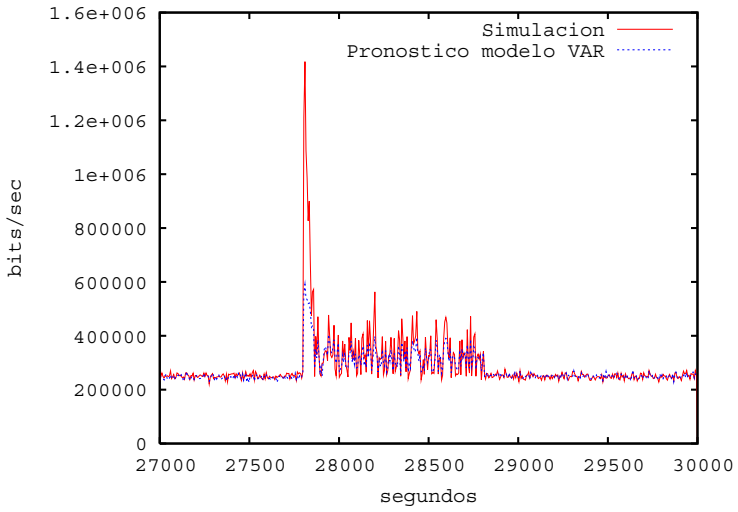

Figura 5.19.: Tráfico enviado del canal ethernet del servidor FTP (bits/sec).

La gráfica 5.20 muestra una comparación de los valores obtenidos del modelo de pronóstico y los que se obtuvieron en la simulación de la variable V_4. Se puede ver en la gráfica que durante el periodo de validación esta variable solo se presenta por un tiempo de alrededor de 1000 segundos. En esta gráfica también se logra pronosticar los cambios abruptos de tráfico. El error(promedio) de pronóstico de esta variable es de 0.8864 %.

La gráfica 5.21 muestra una comparación de los valores obtenidos del modelo de pronóstico y los que se obtuvieron en la simulación de la variable V_5. Se puede ver en la gráfica que durante el periodo de validación esta variable solo se presenta por un tiempo de alrededor de 600 segundos. En esta gráfica también se logra pronosticar los cambios abruptos de tráfico. El error(promedio) de pronóstico de esta variable es de 0.9467 %. Tanto en la variable V_4 como en la V_5, se presentaron buenos resultados de pronóstico.

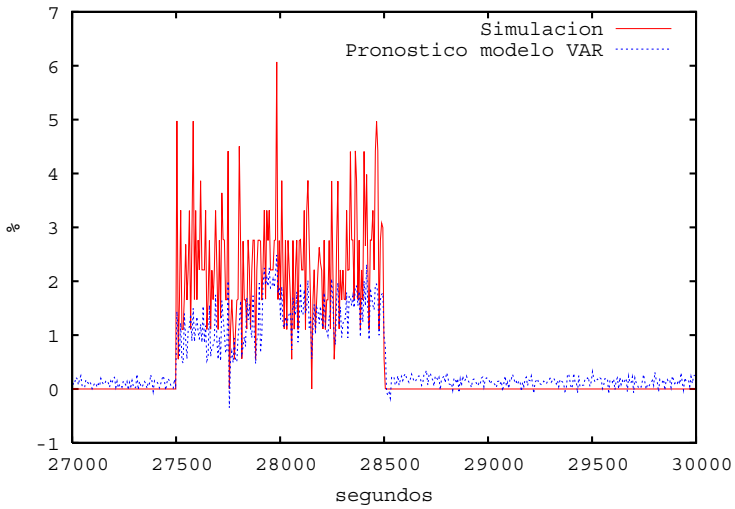

Figura 5.20.: Porcentaje de utilización del CPU del servidor de base de datos (%).

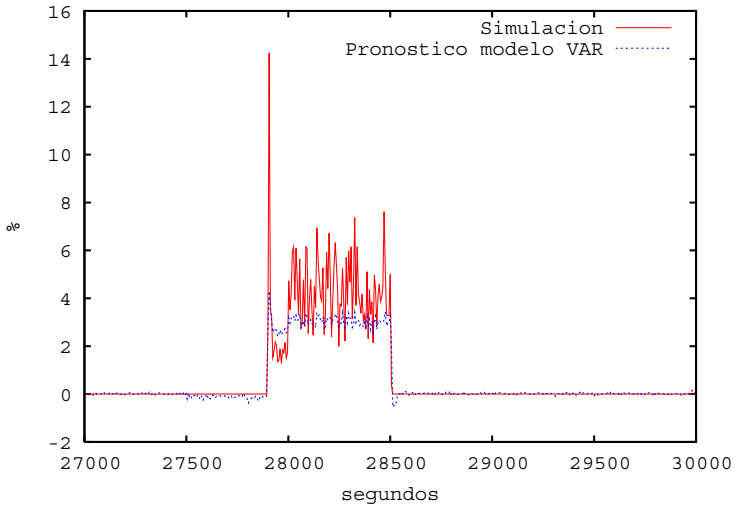

Figura 5.21.: Porcentaje de utilización del CPU del servidor de web (%).

La gráfica 5.22 y , 5.23 presentan el pronóstico de la variable V_6 y V_7, respectivamente. En ambas variables se logra pronosticar los cambios abruptos de tráfico. El error(promedio) de pronóstico de V_6 y V_7 son de 73.101 kbps y 72.324 kbps. Con estos resultados se puede mostrar que se tienen buenos resultados de pronóstico, ya que el tráfico en estos servidores es de hasta 600 kbps, de tal manera que tener un error de pornóstico del 73.101 kbps y 72.324 es aceptable.

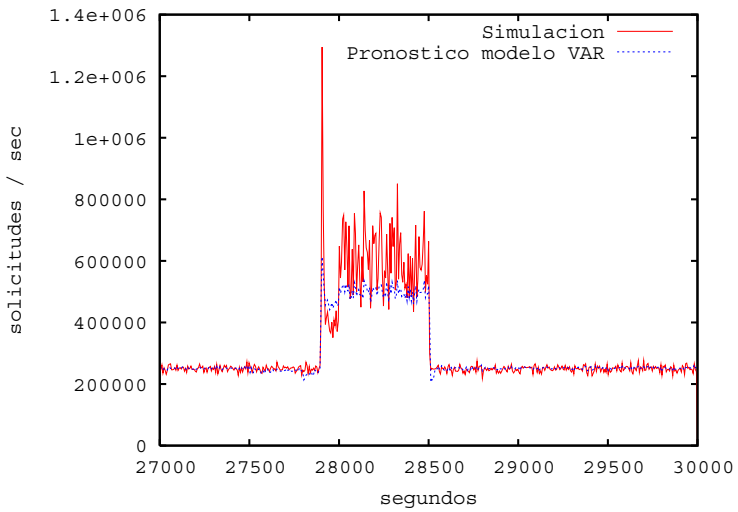

Figura 5.22.: Tráfico enviado del canal ethernet del servidor web (bits/sec).

Para medir realmente la eficiencia del modelo de pronóstico, se necesita realizar la etapa de predicción del tráfico, en donde se medirá el nivel de error promedio. En general se puede observar que en la mayoría de las variables se lograron pronosticar los cambios abruptos de sus valores, lo cual es realmente uno de los objetivos del modelo de pronóstico.

5.6. Predicción del tráfico de la red.

Ahora que tenemos los valores de las 7 entradas en un tiempo t+1, es posible utilizar un modelo de predicción. El modelo de predicción como se explico en el capítulo 4, esta basado en una Red Neuronal MLP, como se muestra en la figura 4.7. La red MLP tiene

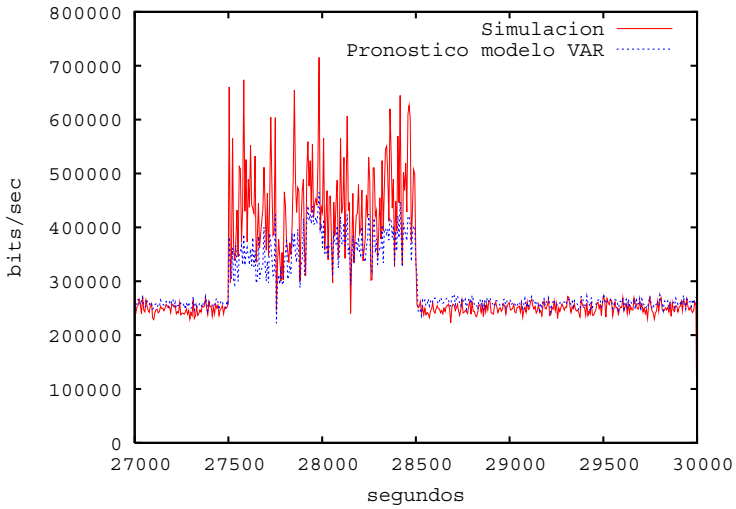

Figura 5.23.: Tráfico enviado del canal ethernet del servidor de base de datos (bits/sec).

7 variables de entrada, una capa oculta con 6 unidades y una sola salida. Con las 7 variables de entrada se puede predecir el valor del target en un tiempo t+1. El modelo de predicción se implementa en SAS Enterprise Miner por medio del nodo de RN.

En la gráfica 5.24 se presentan los valores obtenidos de la simulación y de la predicción de la variable de tráfico. La predicción en la grafica 5.24 es durante el conjunto de datos de validación de la variable tráfico. Con estos resultados se demuestra que en esta etapa de la metodología se obtiene la predicción de la utilización del enlace entre las subredes de la Ciudad de México y Monterrey. Este conocimiento adquirido se puede utilizar para ayudar a la planeación de la capacidad de la red de datos.

La tabla 5.9 presenta la SSM, SME, RSME de la predicción del tráfico. Estos valores son durante el periodo de validación del modelo de predicción.

5.7. Interpretación de los resultados

Como se menciono anteriormente para conocer el desempeño de una red neuronal se mide su función de error durante la etapa de entrenamiento y se selecciona el modelo

Tabla 5.9.: Medidas de desempeño del modelo de predicción.

Variable	**SSE**	**MSE**	**RMSE**
tráfico	22001.013625	44.00202725	6.63340238

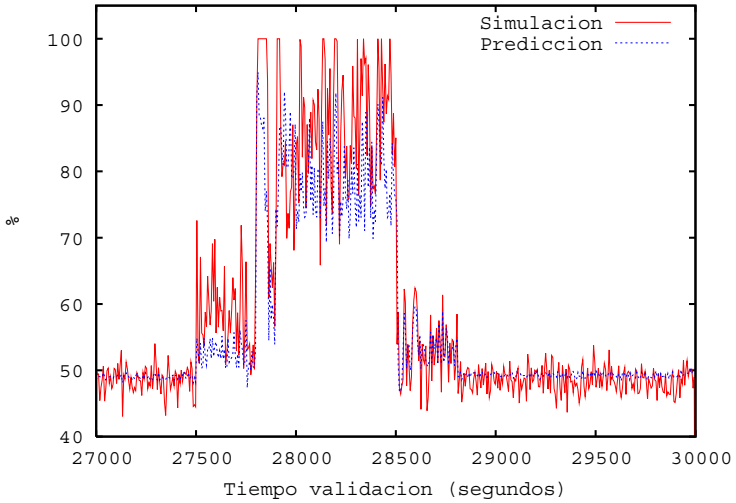

Figura 5.24.: Predicción del porcentaje de utilización del enlace entre las subredes de la Ciudad de México y Monterrey.

que tenga el valor de la función de error más pequeño.

SAS Enterprise Miner proporciona la SSM, SME y RSME durante la etapa de entrenamiento y validación del modelo de predicción. Los valores obtenidos de estas medidas de desempeño en la etapa de validación son: SSE de 22001.013625 %, MSE 44.002 % y RMSE 6.63 %. Tener un error (promedio) de 6.63 % entre el tráfico que se espera y el que se predice es un buen margen de error para predecir el nivel de utilización entre las subredes de la Ciudad de México y Monterrey. Los valores de estas medidas de desempeño comprueban que los resultados que se tienen en la grafica 5.24, presentan un buen porcentaje de predicción.

Capítulo 6.

Conclusiones

Las técnicas de inteligencia analítica han demostrado que pueden solucionar múltiples problemas en redes de telecomunicaciones. De entre todos los modelos que utiliza es la minería de datos la que provee un conjunto de técnicas para desarrollar modelos de pronóstico de series de tiempo y predicción de variables de red.

En esta tesis se plantea una metodología que permite resolver el problema de predicción del tráfico de los enlaces en una red de datos. Esta metodología se aplico a un escenario de red de datos con la finalidad de demostrar que en cada uno de los pasos que la integran se van obteniendo resultados parciales que permiten entender la topología, estructura y aplicaciones de la red de datos, sin tener que ser necesariamente experto en el área de las telecomunicaciones para entender el tráfico de la red.

La metodología propuesta permitió pasar de un conjunto de estadísticas que describen el comportamiento del tráfico, hasta conocimiento que permite predecir el tráfico de la red. Durante este proceso de transformación se seleccionaron diversas técnicas de Inteligencia Analítica. En la transformación de datos a información se utilizo la correlación de Spearman para determinar las variables no redundantes de la red y árboles de decisión para seleccionar las variables más importantes que estuvieran relacionadas con la variable de tráfico. En la transformación de información a conocimiento, se propuso la utilización de un modelo vector autoregresión basado en una red neuronal perceptron para el pronóstico de variables de red y la utilización de MLP para la predicción del tráfico, con los resultados del caso de estudio se demostro que estas técnicas están bien situadas para resolver este problema. Por último la transformación de conocimiento en inteligencia, consiste en poder integrar los resultados obtenidos en un escenario de red real.

La ventaja que presenta la aplicación de la metodología es que el pronóstico de tráfico no lo basamos en una sola variable, realizamos un proceso de selección de variables

de red que esten relacionadas con la predicción del tráfico. El análisis se realizo para detectar aquellas variables que fueran viables de obtener de un escenario de red real.

Hay varias líneas de investigación que se podrán seguir a partir de este trabajo de investigación:

- La más importante consiste en poder integrar esta metodología a un sistema de tiempo real, en donde todo el tiempo se estén monitoreando las variables de red seleccionadas en la metodología, y que ya sea por implementaciones de hardware o software se pueda estar informando la predicción del tráfico en los enlaces en la red.
- Otra investigación consistiría en poder integrar los valores de predicción a un sistema inteligente que permita responder de manera automática cuando se hagan predicciones indicando niveles de tráfico muy alto.
- Una etapa que se puede mejorar durante el desarrollo de la metodología, es mejorar los modelos de pronóstico y predicción. Se podría plantear un modelo de pronóstico basado en redes MLP y para el modelo de predicción tratar de combinar algunas técnicas de selección para disminuir la función de error durante la etapa de validación de los datos.

Apéndice A.

Simulación en Opnet Modeler

Opnet Modeler brinda un ambiente de desarrollo que permite diseñar y estudiar redes, dispositivos, protocolos y aplicaciones de comunicaciones con un grado de flexibilidad. Opnet ha ayudado a la industria de las telecomunicaciones a planear las capacidades de las redes de telecomunicaciones. Dentro de las múltiples tareas que permite realizar Opnet Modeler estan se encuentran:

- Simular redes de comunicaciones del mundo real.
- Coleccionar estadísticas sobre el desempeño de la red.
- Analizar las estadísticas de simulación.
- Configurar una paleta de objetos de acuerdo a los modelos que se necesitan.
- Configurar de aplicaciones y perfiles.
- Modelar redes de área local (LAN).
- Especificar cambios en el tiempo de los niveles de utilización en los enlaces.
- Simular múltiples escenarios simultáneamente.
- Generar reportes web y usar esta información para tomar decisiones sobre el comportamiento de la red.
- Importar tráfico.

Las librerías de Opnet Modeler son reconocidas en la industria de las telecomunicaciones dentro del conjunto mas avanzado de modelos, protocolos, tecnologías y aplicaciones que se tienen disponibles.

La figura A.1, presenta un ejemplo de un escenario de red de una universidad con siete redes de área local (LAN) conectadas a un backbone ATM. La red tiene redes

LAN Ethernet, FDDI y Token Ring cada una con un determinado número de clientes y con una determinada configuración de aplicaciones. Este escenario simulado tiene carga FTP baja. Cada red LAN soporta todas las aplicaciones excepto la de FTP, la cual es soportada solo por el servidor FTP.

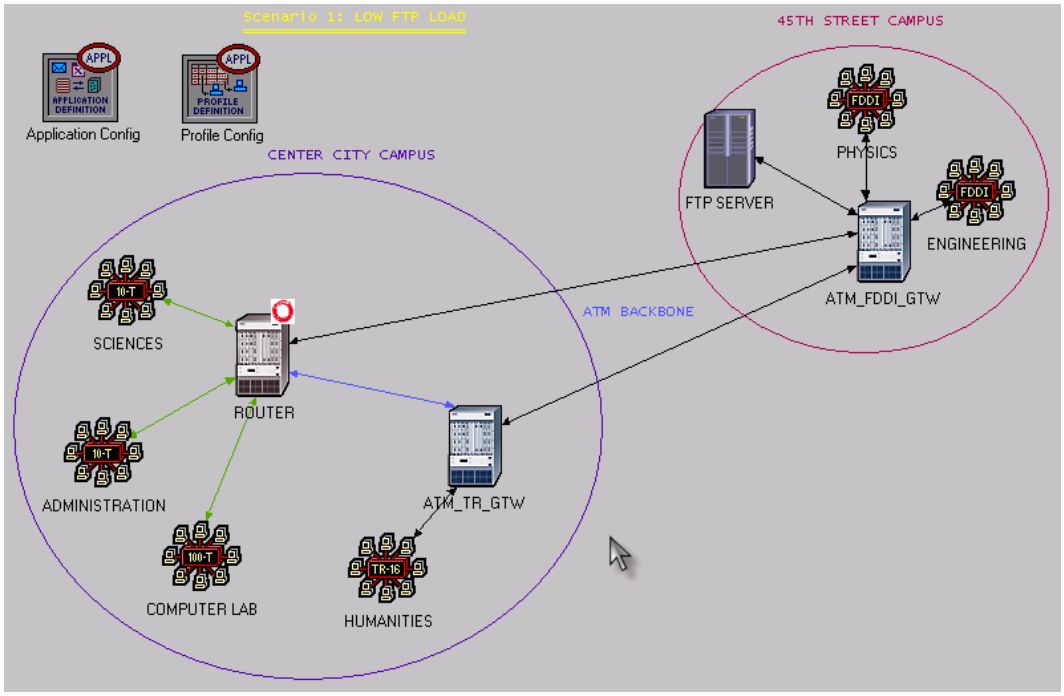

Figura A.1.: Escenario de una red WAN.

Opnet permite definir distintas aplicaciones: Acceso a base de datos, Email, transferencia de archivos, impresión de archivos, sesiones telnet, video conferencia, voz sobre IP y web browsing. Cada una de estas aplicaciones se pueden configurar para que tengan diferentes niveles de carga: bajo, medio o alto.

A partir de este conjunto de aplicaciones se pueden configurar perfiles de acuerdo a los tipos de servicios requeridos por cada usuario. En cada perfil se tienen que establecer los siguientes parámetros:

- *Selección de las aplicaciones que tendrá este perfil.* Por cada aplicación se podrá definir el tiempo de inicio, duración y periodo de repetición. Estos tiempos se tienen que definir durante el periodo de duración del perfil, aquéllos que sobrepasen la duración del perfil no se ejecutarán.

- *Modo de operación.* Consiste en especificar el orden en que se irán ejecutando los perfiles. Puede ser serial, que consiste en ejecutar solo un perfil a la vez, o simultaneo que consiste en ejecutar todos los perfiles al mismo tiempo.
- *Tiempo de inicio.* Consiste en especificar durante todo el tiempo que dura la simulación el tiempo en que iniciara el perfil.
- *Duración.* Es el tiempo que dura el perfil.
- *Repetición.* Es el periodo con el que se estará ejecutando el perfil.

La figura A.2 muestra la configuración del perfil del escenario de red de la figura A.1.

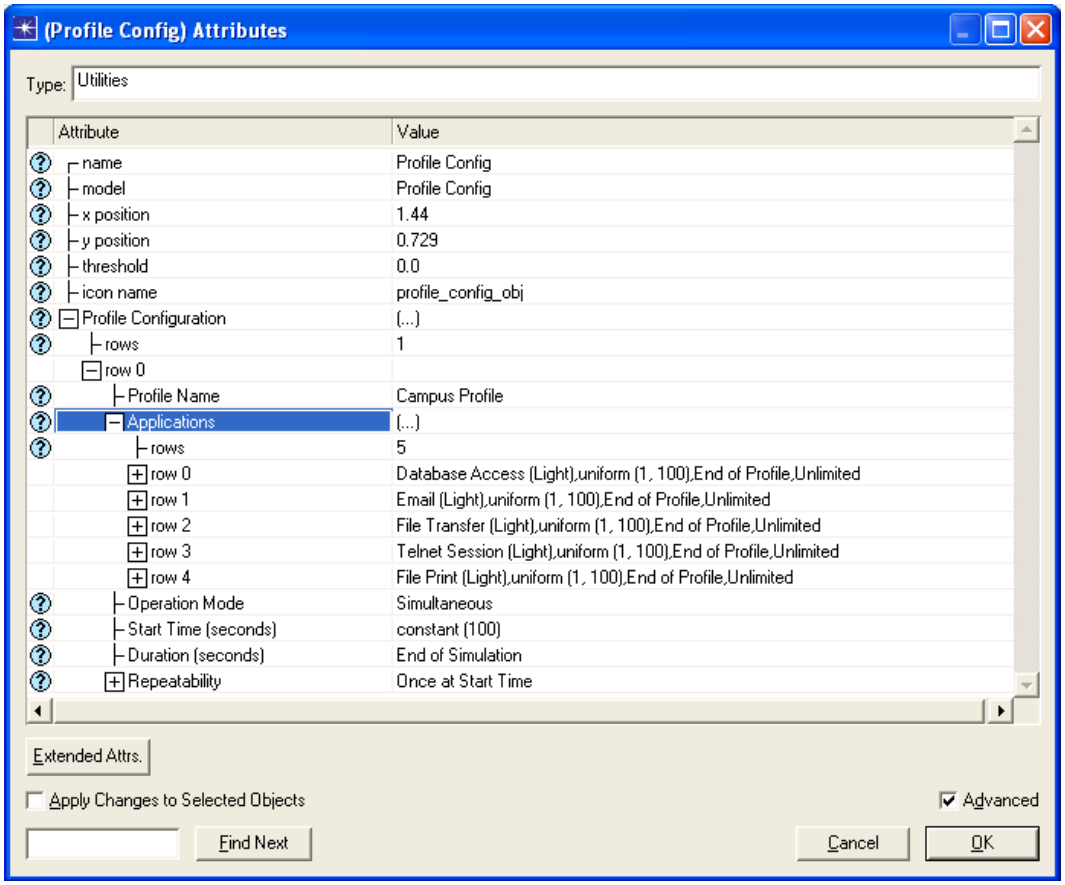

Figura A.2.: Perfil de todas las redes de área local (LAN).

Opnet organiza las estadísticas de la red en tres grupos: estadísticas globales, por nodo y por enlace. La selección del tipo de estadísticas depende de los objetivos de

cada simulación.

En el ejemplo de la figura A.1. Se seleccionaron las siguientes estadísticas:

Tipo de estadística	**Estadísticas**
Generales	ATM, email, ftp, print, remote login.
Por nodo	Client BD, client ftp, client print, client email, server email, server ftp, server print y server remote login.

Tabla A.1.: Estadísticas seleccionadas del escenario de red de la figura A.1

Hay dos propiedades importantes que se tienen que definir en la simulación: la duración y en número de valores por estadística. La duración de la simulación esta limitada por el número de eventos que se pueden utilizar, y el número de eventos se incrementa de acuerdo a la cantidad de tráfico que se maneje en la red, la figura A.3 muestra la configuración de estos parametros, donde la duración de la simulación es de 1800 segundos y 100 valores por estadística. Estos valores indican que las estadísticas que se obtengan de la simulación van a tener 100 muestras, espaciadas 18 segundos.

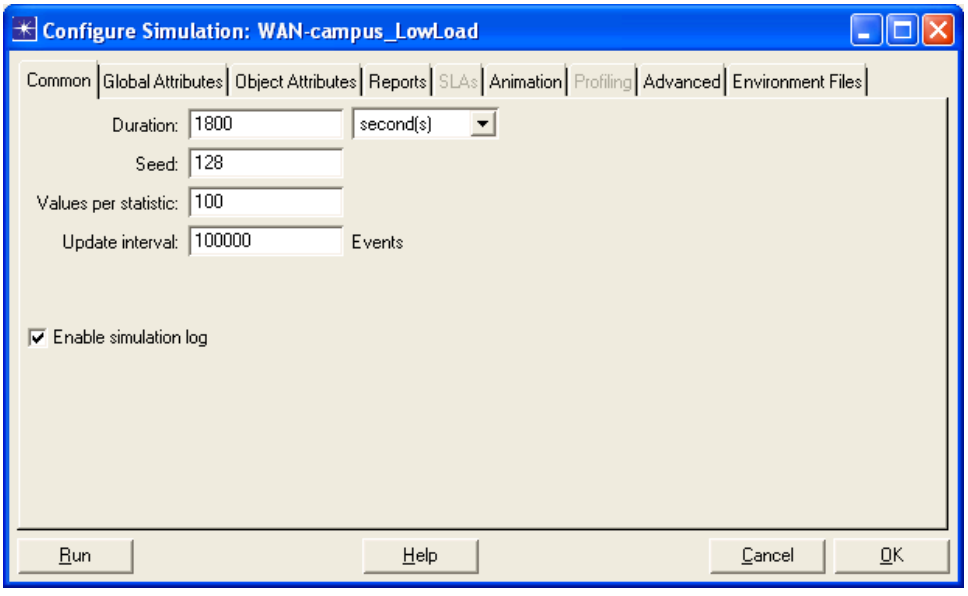

Figura A.3.: Configuración de los parametros de la simulación.

Como resultado de la simulación obtenemos valores de las estadísticas seleccionadas. La figura A.4, muestra la estadística del tráfico enviado por el servidor FTP (bytes/sec).

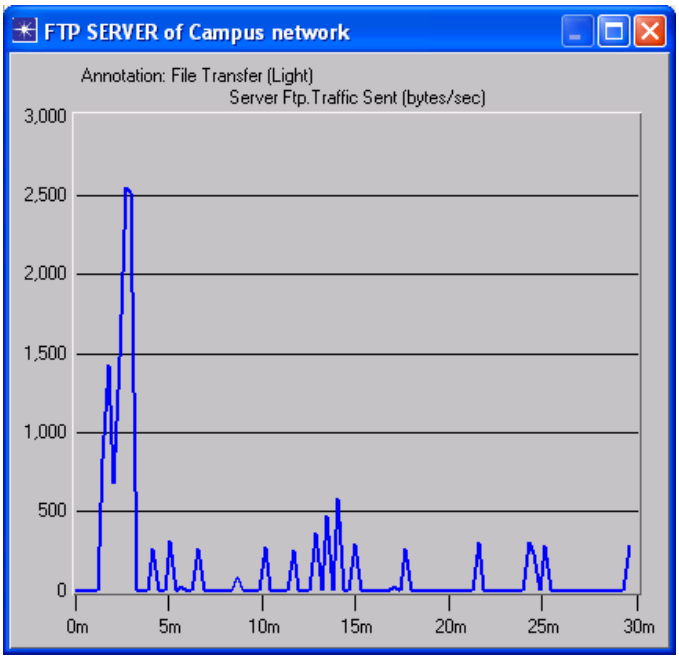

Figura A.4.: Estadística del Tráfico enviado por el servidor FTP.

Apéndice B.

Limpieza de los datos

```
#include <stdio.h>
#include<conio.h>
#include<string.h>
#include<stdlib.h>

int main(void) {
   FILE *in, *out;
   int i=0,tmp=0,j=0;
   char *string;
   char link[45];
   char limpio[45];

   strcpy(link,"\\data set\\");
   strcpy(limpio,"\\data set\\sinruido\\");
   printf("\n Nombre archivo:  ");
   gets(string);
   strcat(link, string);
   strcat(limpio, string);

   if ((in = fopen(link, "r")) == NULL)
   {  fprintf(stderr, "Nombre Incorrecto.\n");
      return 1;
   }
   if ((out = fopen(limpio, "w")) == NULL)
   {  fprintf(stderr, "No se creo archivo de correciones.\n");
      return 1;
   }
   while (!feof(in)){
```

```
      tmp = fgetc(in);
      if(tmp == 35){ j=0;
            while( j < 3){
                tmp = fgetc(in);
                j++;
            }
            tmp='0';
            fputc(tmp, out);
            tmp='.';
            fputc(tmp, out);
            tmp='0';
            for(j=0;j<9;j++)
                fputc(tmp, out);
      }
      else
            if(!feof(in))fputc(tmp, out);
   }
   fclose(in);
   fclose(out);
   return 0;
}
```

Apéndice C.

Análisis de Correlación

```
#include<stdio.h>
#include<conio.h>
#include<string.h>
#include<math.h>
#include<stdlib.h>
#define N 500
#define T 191

int main(void) {

float U=0.8;
int vars[N]={0};
int i=0,j=0,k=0,z=0, e=0;
char *string, char link[45], charlimpio[45];
int tmp, var=0,cont=0,cont2=1, vacio=0;
float valor=0.0;
int ctarget=0, corr=0,laste=0, espacio=0;
char copytarget[500]={''};
float valorlog=0.0;
FILE *in,*out,*cop; int
eliminados[T]={2,3,4,5,37,38,39,67,68,69,92,93,94,115,116,117,150,151,
               152,153,154,155};

   strcpy(link,"\\data set\\matriz\\");
   strcpy(limpio,"\\data set\\matcor\\");
   printf("\n Nombre archivo:  ");
   gets(string);
   strcat(link, string);
```

```
   strcat(limpio, string);

   if ((in = fopen(link, "r")) == NULL)
   {    fprintf(stderr, "No se abrio el archivo.\n");
        return 1;
   }
   if ((out = fopen(limpio, "w+"))== NULL) {
        fprintf(stderr, "Cannot open output file.\n");
        return 1;
    }

/* Contar el número de variables */

   tmp = fgetc(in);
   while(tmp != '\n'){
        if(tmp == '\t' ) var++;
        tmp = fgetc(in);
   }

   for(k=0;k<var;k++)
        vars[k]=k+1;

   for(k=1,e=0;k<var;k++){
        if(eliminados[e]==k){
            vars[k-1] = 0;
            e++;
        }
    }

    for(k=153;k<var;k++)
        vars[k-1] = 0;

    while (!feof(in)){
        tmp = fgetc(in);
        while(tmp != '\t')
            tmp = fgetc(in);
        cont=0;
        j=0;
        vacio=0;
```

```
        while(cont<var){
            fscanf(in,"%f ",&valor);
            if(valor == 0) vacio ++;
            if( valor > U )
            if(vars[i] != 0 && vars[j] != 0 && i!=j)
            vars[j]=0;
            valor=0.0;
            j++;
            cont++;
        }
        if(vacio==var) vars[i]=0;
        i++;
        if(cont2==100) getch();
    }

/* Numero de la variable de trafico en el conjunto de datos */

    vars[168]= 169 ;

    for(cont2=0,k=0;k<var;k++){
        printf("%d\t",vars[k]);
        if(vars[k] != 0){ cont2++; laste = k+1;}
    }

/* Si la correlacion es mayor a U, solo se considera a una variable
   Si se elimina una variable indica que hay otra variable que la
   representa porque tienen alta correlacion, esa variable ya no se
   vuelve a considerar. */

    if (( cop= fopen("\\data set\\sinruido\\dt_191.txt", "r"))== NULL) {
        fprintf(stderr, "No se puede abrir el archivo.\n");
        return 1;
    }

    var=1;
    fseek(cop,0, SEEK_SET);
    tmp = 0;
    cont=0;
```

```
while(tmp != 10){
    tmp = fgetc(cop);
    if(var == vars[cont] && tmp != 10 ){
        z=1;
        copytarget[ctarget++]=tmp;  /* Guarda nombre del target */
        fputc(tmp, out);
    }
    if(tmp == 9 || tmp == 10){
        cont++;
        if(z == 1){
            z=0;
            if(var != 1){
                for(i=0;i<ctarget;i++){
                    if(cont == laste && i == ctarget-1);
                    else
                        fputc(copytarget[i],out);
                }
            }
        ctarget=0;
    }
    var++;
    }
}
var--;
corr=0;
espacio=0;
cont=1;

fputc('\n', out);

while (!feof(cop)){
    while(espacio < var && !feof(cop)){
        fscanf(cop,"%f ",&valor);
        espacio++;
        if(espacio == vars[corr]){
            valorlog=valor;
            if(valor != 0 && espacio != 1) valorlog= log(valor);
            if(espacio != 1)
                fprintf(out, "%f\t",valorlog);
```

```
            if(corr == laste-1)
                fprintf(out, "%f",valor);
            else
                fprintf(out, "%f\t",valor);
        }
    corr++;
    }
fputc('\n',out);
espacio=0;
corr=0;
}

fclose(in);
fclose(out);
fclose(cop);
return 0; }
```

Bibliografía

[1] Inteligencia analítica, sas. http://www.sas.com.

[2] Network capacity planning, ncr. http://www.ncr.com.

[3] Opnet technologies. opnet modeler 9.1 http://www.opnet.com.

[4] Optimizing network traffic, http://www.microsoft.com.

[5] N. Sadek A. Khotanzad. Multi-scale high-speed network traffic prediction using combination of neural networks. *Proceedings of the International Joint Conference on Neural Networks*, pages 1071–1075, 2003.

[6] Nagpal S. Sundararajan N. Saratchandran P. Aiyar, M. Minimal resource allocation network (mran) for call admission control (cac) of atm networks. *Proceedings. IEEE International Conference*, 2000.

[7] Olshen RA Stone CJ Breiman L, Friedman JH. *Classification and Regression Trees*. New York, 1984.

[8] A. Brischetto and G. Voss. A structural vector autoregression model of monetary policy in australia. *Economic Research Department, Reserve Bank of Australia*, 1999.

[9] E. Pednault B. Rosen F. Tipu C. Apte, E. Grossman and B. White. Probabilistic estimation based data mining for discovering insurance risks. *IBM Research Division*, 1999.

[10] D. Wei C. Guang, G. Jian. Nonlinear-periodical network traffic behavioral forecast based on seasonal neural network model. *International conference on communications, circuits and systems.*, pages 683–687, 2004.

[11] Sims C.A. Macroeconomics and reality. *Econometrica*, 1980.

[12] Vincent Wing-Sing Cho. Knowledge discovery from distributed and textual data. *PhD thesis, Hong Kong University of Science and Technology*, 1999.

[13] D. H. Diez de Medina. Una estimación del pib potencial basada en restricciones de corto plazo. *Unidad de Análisis de Políticas Sociales y Económicas*, 2005.

[14] S. Inglis G. Holmes Ian H. Witten E. Frank, Y. Wang. Using model trees for classification. *Department of Computer Science, University of Waikato*, 1997.

[15] Jesus Rico Edgar N. Sanchez, Alma Y. Alanis. Predicción de la demanda eléctrica usando redes neuronales entrenadas por filtro de kalman. *XI Congreso Latinoamericano de Control Automático*, 2004.

[16] Ming Wu Eswaradass, A. Xian-He Sun. Network bandwidth predictor (nbp): A system for online network performance forecasting. *Cluster Computing and the Grid*, pages 265–268, 2006.

[17] H. Ding G. Tan, W. Yuan. Traffic flow prediction based on generalized neural network. *IEEE lnlelllgenl Transpollation Systems Conference*, 2004.

[18] J. Del Carpio Gallegos. Las redes neuronales artificiales en las finanzas. *Revista de la Facultad de Ingeniería Industrial Vol. (8)*, pages 28–32, 2005.

[19] M. Garofalakis and R. Rastogi. Data mining meets network management: The nemesis project. *ACM SIGMOD Int'l Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2001.

[20] Rob Gerritsen. Assessing loan risks: A data mining case study. *IT Professional*, pages 16–21, 1999.

[21] H. Xiao B. Ran H. Sun, Henry X. Liu. Short term traffic forecasting using the local linear regression model. *Institute of Transportation Studies University of California,*, 2002.

[22] J. Hall and P. Mars. Limitation of artificial neural networks for traffic prediction in broadband networks. *IEE Proceedings: Communications*, pages 114–118, 2000.

[23] Fred Halsall. *Data Communications, Computer Networks and Open Systems*. Adison-Wesley Publishing Company, 1996.

[24] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, British Columbia, Canada, 2003.

[25] Klemettinen M. Mannila H. Ronkainen P. Hätönen, K. and H. Toivonen. Knowledge discovery from telecommunication network alarm databases. 1996.

[26] Z. Zhang K. Papagiannaki, N. Taft and C. Diot. Long-term forecasting of internet backbone traffic: Observations and initial models. *In IEEE INFOCOM*, 2003.

[27] Mika Klemettinen. A knowledge discovery methodology for telecommunication network alarm databases. *University of Helsinki Finland*, 1999.

[28] Dominic Savio Lam Lai Yin. Learned text categorization by backpropagation neural network. *The Hong Kong University of Science and Technology*, 1996.

[29] Faming Liang. Bayesian neural networks for nonlinear time series forecasting. *Statistics and Computing 15*, 2004.

[30] J. Shavlik M. Craven. Using neural networks for data mining. *Future Generation Computer Systems*, 1997.

[31] A. Hussain M. Jaudet, N. Iqbal. Neural networks for fault-prediction in a telecommunications network. *Proceedings of INMIC*, pages 315– 320, 2004.

[32] E. Raftopoulos M. Ploumidis F. Hernandez-Campos M. Papadopouli, H. Shen. Short-term traffic forecasting in a campus-wide wireless network. *16th Annual IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, 2005.

[33] I. G. Martín. Análisis y predicción de la serie de tiempo del precio externo del café colombiano utilizando redes neuronales artificiales. *Revista de la Facultad de Ciencias Pontificia Universidad Javeriana Vol. (8)*, pages 45–50, 2003.

[34] Gordon S. Linoff Michel J.A. Berry. *Data Mining Techniques*. Wiley Publishing, Inc., Indianapolis, Indiana, 2004.

[35] Kolluru Venkata Sreerama Murthy. On growing better decision trees from data. *Ph.D. dissertation, Univ. of Maryland, College Park*, 1997.

[36] I. Miloucheva. N. Toumba. Pattern based spatio-temporal quality of service analysis for capacity planning. *First international workshop on Inter-domain performance and simulation*, 2003.

[37] T. Funabashi P. Mandal, T. Senjyu. Forecasting several-hours-ahead electricity demand using neural network. *Electric Utility Deregulation, Restructuring and Power Technologies*, pages 515– 521, 2004.

[38] N. Davey y S.P. Hunt. R.J. Frank. Time series prediction and neural networks. *Intelligent and Robotic Systems*, pages 91–103, 2001.

[39] Kattamuri S. Sarma. Using sas enterprise miner for forecasting. *SAS Users Group International (SUGI 26)*, pages 111–115, 2001.

[40] Kattamuri S. Sarma. Combining decision trees with regression in predictive modeling with sas enterprise miner. *SAS Users Group International (SUGI 30)*, 2005.

[41] Y. Shachmurove and D. Witkowska. Utilizing artificial neural network model to predict stock markets. *Department of Economics, The City College of the City University of New York*, 2000.

[42] Shapcott C.M. y Curran E.P. Sterritt R., Adamson K. Data mining telecommunications network data for fault management and development testing. *IEE Proc.-Comm.*, pages 299–308, 2000.

[43] R. Stahlbock Sven F. Crone, S. Lessmann. Utility based data mining for time series analysis cost-sensitive learning for neural network predictors. *UBDM'05*, 2005.

[44] R. J. Frank N. Davey T. Edwards, D.S.W. Tansley. Traffic trends analysis using neural networks. *Proceedings of the International Workshop on Applications of Neural Networks to Telecommuncations*, pages 157–164, 1997.

[45] Gary. M. Weiss. Intelligent telecommunication technologies. *Knowledge-Based Intelligent Techniques in Industry (chapter 8)*, pages 249–275, 1998.

[46] Gary. M. Weiss. Predicting telecommunication equipment failures from sequences of network alarms. *In Handbook of Knowledge Discovery and Data Mining*, 2001.

[47] Gary M. Weiss. The effect of small disjuncts and class distribution on decision tree learning. *PhD thesis, School- New Brunswick Rutgers, The State University of New Jersey*, 2003.

[48] Gary M. Weiss. Data mining in telecommunications. *Departament of Computer and Information Science*, 2005.

[49] IndurkhyaÑ. Weiss S. Predictive data mining: A practical guide. Technical report, San Francisco, USA, 1998.

[50] Eilyn Arias C. y Carlos Torres G. Modelos var y vecm para el pronóstico de corto plazo de las importaciones de costa rica. *Departamento de Investigaciones Economicas, Banco Central de Costa Rica*, 2004.

[51] C.N. Manikopoulos J. Jorgenson J. Ucles Z. Zhang, J. Li. Hide: a hierarchical network intrusion detection system using statistical preprocessing and neural network classification. *Proceedings of the 2001 IEEE, Workshop on Information Assurance and Security*, 2001.