Minería de Datos aplicada a la Detección de factores para la prevención de incidentes informáticos.

Corso Cynthia, García Alejandro, Ciceri Leonardo, Romero Fernando.

Laboratorio de Sistemas de Información
Departamento de Ingeniería en Sistemas de Información
Facultad Regional Córdoba
Universidad Tecnológica Nacional.
Maestro M. López esq. Cruz Roja.
cynthia@bbs.frc.utn.edu.ar/malejandrogarcia@hotmail.com
/leocic@bbs.frc.utn.edu.ar/48933@sistemas.frc.utn.edu.ar

Resumen

El objetivo principal de este proyecto es la aplicación de técnicas de minería de datos para lograr la caracterización de incidentes informáticos de los recursos tecnológicos, en el contexto de un laboratorio informático.

Palabras clave: Minería de Datos, Incidente Informático, Laboratorio Informático, Weka.

Contexto

Este trabajo presenta el resultado de los avances del proyecto de investigación "Generación de Modelo Descriptivo para la prevención de incidentes de equipos informáticos en el contexto de laboratorio de sistemas", cuyo código es UTN 1683 homologado por la Secretaría de Ciencia, Tecnología y Posgrado de la Universidad Tecnológica Nacional.
El contexto de trabajo es un Laboratorio Informático con fines meramente académicos y de investigación, que depende del Departamento de Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional Facultad Regional Córdoba.

Introducción

Uno de los principales problemas en el mundo laboral actual es el elevado número de incidentes que se producen en los equipos laborales. Toda organización debe tener el total conocimiento y puesta en práctica de medidas de prevención que garanticen el normal funcionamiento de los mismos.

[1] El análisis de los registros de los reportes de incidentes y mantenimientos es fundamental en la prevención de incidentes informáticos. Entre otras cosas, porque permite el diseño de políticas y planes de prevención efectivos.

Las técnicas que ofrece la estadística son innumerables y permiten realizar un

análisis más que exhaustivo. Sin embargo, en muchas ocasiones la estadística descriptiva clásica no refleja la verdadera interrelación de las variables y por lo tanto, el problema real.

[2] En la actualidad el tamaño de las bases de datos está basado en aspectos como la capacidad y eficiencia de almacenamiento y no en su posterior uso o análisis. Por esta razón, en muchos casos, los registros almacenados son demasiado grandes o complejos como para analizar y superan el alcance de la estadística. Frente a este contexto resulta necesaria una alternativa superadora que permita cubrir las limitaciones de la estadística y facilite la caracterización de incidentes informáticos.
Una opción considerada como atrayente e interesante para este estudio es la Minería de Datos.

[3] La minería de datos se define como el proceso de exploración y análisis, por medios automáticos o semiautomáticos, de grandes volúmenes de información con el objetivo de descubrir e identificar patrones y reglas significativas.

El objetivo principal de la minería de datos es el de analizar los datos para extraer conocimiento, este puede encontrarse en forma de relaciones, patrones o reglas, que precisamente serán inferidas de los datos, o bien en forma de una descripción mas concisa. La minería de datos tiene una serie de tareas que pueden interpretarse como un tipo de problema a ser resuelto por un algoritmo de minería de datos. Esto significa que cada tarea tiene sus propios requisitos, y que el tipo de información obtenida con una tarea puede diferir mucho de la obtenida con otra.

[4] Los modelos pueden ser de dos tipos: Predictivos y Descriptivos. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo o dependientes, usando otras variables o campos de las bases de datos que se denominan variables independientes o predictivas.

Los modelos descriptivos identifican patrones que explican o resumen los datos, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos.
En este estudio se implementan las siguientes técnicas de minería de datos.

[5] Técnicas de Agrupamiento: El Análisis de Clusters (o Análisis de conglomerados) es una técnica de Análisis Exploratorio de Datos para resolver problemas de clasificación. Su objeto consiste en ordenar objetos (personas, cosas, animales, plantas, variables, etc.) en grupos (conglomerados o clusters) de forma que el grado de asociación/similitud entre miembros del mismo clúster sea más fuerte que el grado de asociación/similitud entre miembros de diferentes clusters.

Cada clúster se describe como la clase a la que sus miembros pertenecen. El agrupamiento puede realizarse tanto para casos como para variables, pudiéndose utilizar variables cualitativas o cuantitativas. El análisis de clúster es un método que permite descubrir asociaciones y estructuras en los datos que no son evidentes a priori pero que pueden ser útiles una vez que se han encontrado. Los resultados de un Análisis de Clusters pueden contribuir a la definición formal de un esquema de clasificación tal como una taxonomía para un conjunto de objetos, a sugerir modelos

estadísticos para describir poblaciones, a asignar nuevos individuos a las clases para diagnóstico e identificación, etc.

[6] Técnicas de Asociación: Las reglas de asociación son una poderosa técnica de minería de datos, y son utilizadas para buscar por medio de conjuntos de datos reglas que revelan la naturaleza de las relaciones o asociaciones entre datos de las entidades. Las asociaciones resultantes pueden ser utilizadas para filtrar la información, para analizarlas y posiblemente para definir un modelo de predicción basado en la observación del comportamiento.

El análisis de asociación, que persigue el establecimiento de relaciones entre registros individuales o grupos de registros de la bases de datos. Una de las especializaciones es el análisis de asociación son reglas de asociación. Las reglas de asociación se utilizan para descubrir relaciones entre los grupos de registros dentro de una base de datos.
Una regla de asociación tiene la forma:

"Si X entonces Y"

Donde:
X: se denomina antecedente de la regla.
Y: consecuente de la regla.

En la práctica, una regla necesita un soporte de varios cientos de registros antes de que ésta pueda considerarse significativa desde un punto de vista estadístico.
A menudo las bases de datos contienen miles o incluso millones de registros. Para seleccionar reglas interesantes del conjunto de todas las reglas posibles que se pueden derivar de un conjunto de datos se pueden utilizar restricciones sobre diversas medidas de "significancia" e "interés". Las restricciones más conocidas son los umbrales mínimos de soporte y confianza.

Líneas de Investigación, Desarrollo e Innovación

- Herramientas de análisis exploratorio de datos
- Técnicas de Análisis Multivariante.
- Técnicas para la detección y tratamiento de valores atípicos.
- Técnicas de imputación de datos para valores ausentes.
- Técnicas de reducción de datos.
- Algoritmos de agrupamiento y de asociación.
- Herramientas de visualización de resultados
- Metodología para el tratamiento de incidentes en el contexto de equipos informáticos.

Resultados y Objetivos

El problema central que intenta resolver esta investigación es mitigar los tiempos de inactividad de los equipos informáticos del laboratorio. Esta situación se presenta fundamentalmente por el reporte de incidentes informáticos, lo que genera el aislamiento del equipo de aula donde habitualmente funciona. Sumado a los tiempos generados por tareas de mantenimiento correctivo.

Dentro de las etapas del KDD (Knowledge Discovery in Database.) se han realizado los siguientes avances:

Preparación de datos

Fase de Selección: Las tareas cumplimentadas en esta fase se detallan a continuación:

a. Determinación del conjunto bajo estudio, que en este caso asciende a las 800 instancias.
b. Detección y selección de los atributos más significativos pertenecientes a los incidentes informáticos.
c. Definición y caracterización de las variables seleccionadas.
d. Diseño e implantación del esquema de almacén de datos.
e. Implementación de cubo OLAP (MOLAP).

Fase de Exploración: Se han estudiado y caracterizado todas las variables bajo estudio, haciendo foco en parámetros como distribución, y simetría. El software utilizado en esta fase fue Weka. Alguno de los resultados obtenidos:

- La gran mayoría de los incidentes informáticos reportados se concentran en el turno mañana.
- Un porcentaje importante de incidentes informáticos refieren a solo tres tipos de fallas, la gran mayoría corresponden a dispositivos de entrada/salida.
- Los mantenimientos de incidentes fueron resueltos en primera instancia, simplemente realizando tareas de limpieza de los componentes.
- Se ha detectado una serie de incidentes que no son "reales". Es decir que se han reportado como tal, pero en el momento de efectuar mantenimiento se ha diagnosticado que su funcionamiento es normal. Se prevé la caracterización de este grupo para determinar que variables son las que tienen mayor incidencia en estos hechos.

Fase de Limpieza de Datos: con el objetivo de mantener consistencia en los datos, se ha efectuado un análisis de gráficos pertenecientes a la fase de exploración (Diagrama de caja y bigotes) detectando en algunas de las variables bajo estudio valores outliers (valores atípicos). Actualmente se está trabajando en técnicas para su tratamiento y mitigar su impacto en la aplicación de las futuras técnicas de Minería de Datos.

Fase de Transformación de Datos:
La gran mayoría de los atributos seleccionados son de tipo nominal, lo cual obligó a efectuar un proceso de categorización de las mismas para facilitar el procesamiento posterior.

A continuación se detalla la futura metodología a implementar en esta investigación a corto/mediano plazo.

- Aplicación de clustering jerárquico de los datos relevantes referidos a los incidentes informáticos. Esto nos permitirá determinar cuál es el número de grupos a considerar y los centroides iniciales.
- Aplicación de clustering no jerárquico, esto permitirá maximizar la homogeneidad dentro de los grupos y heterogeneidad entre los grupos.
- Análisis de los clúster obtenidos y validación con el usuario.
- Aplicación de algoritmos de asociación y clasificación, con el propósito de encontrar definiciones descriptivas que subyace a la pertenencia de los mismos.

Formación de Recursos Humanos

Los integrantes de este proyecto, en su gran mayoría están conformados por docentes pertenecientes al plantel

académico de la carrera de Ingeniería en Sistemas de Información.

Como así también integrantes del equipo de trabajo asesoran en tesis de magister y trabajo de especialidad vinculada con la temática de investigación de este proyecto.

Todos los integrantes docentes del PID han participado del proceso de categorizaciones en investigación dentro del Programa de Incentivos del MECyT; así como en la categorización interna que posee la U.T.N. Además colaboran en el proyecto un becario graduado y alumno.

Referencias

[1] Perversi, Valenga, F., Fernández, E. 2007. "Identificación y Detección de patrones delictivos basada en Minería de Datos". IX Workshop de Investigadores en Ciencias de la Computación. Pág. 385-389.
[2] Kantardzic, M. 2002. "Data Mining: Concepts, models, methods and algorithms." Wiley-IEEE Press.
[3] Berry, M., & Linnof, G. 1999. Mastering data mining: the art and science of customer relationship management.
[4] José, F. 2007. Introducción a la Minería de Datos. Madrid: Pearson.
[5] Araujo, B. S. (2006). Aprendizaje Automático: Conceptos Básicos y Avanzados: Aspectos prácticos usando el software Weka. Prentice-Hall.
[6] Orallo Hernández, J., Ramírez, J., & Ferri, C. 2004. Introducción a la Minería de Datos.
[7] Witten Ian, Frank Eibe, Hall Mark. 2011. Data Mining. Practical Machine Learning Tools and Techniques.3era Edición.ISBN: 978-0-12-374846-0.
[8] Orallo Hernández, J., Ramírez, J., & Ferri, C. 2004. Introducción a la Minería de Datos.
[9] Suárez, M. M. (2006). Análisis Multivariante: Clasificación, Organización y Validación de Resultados. Fourth LACCEI International Latin American and Caribbean Conference for Engineering and Technology. Mayagüez.