WorkShop
EIG2009

WORKSHOP INTERNACIONAL
EIG2009
Departamento de Ingeniería de Sistemas
Universidad de La Frontera – Chile
Diciembre 3 y 4, 2009

Aplicación de Minería de Datos para la Detección de Anomalías: Un Caso de Estudio

Ania Cravero Leal, Samuel Sepúlveda Cuevas
Depto. Ingeniería de Sistemas
Universidad de la Frontera, Temuco, Chile
{acravero, ssepulve}@ufro.cl

Resumen

La Minería de Datos es una de las soluciones de la Inteligencia de Negocios, que ayuda a extraer conocimiento a partir de datos que las empresas han generado producto de su negocio. Este conocimiento puede generar aplicaciones de alto valor agregado si el proceso de Minería de Datos es entendido apropiadamente desde una perspectiva del negocio. Aplicaciones tales como detección de anomalías para la prevención de fraudes y abusos, análisis de fidelización, cross-selling, optimización de la cadena de suministro; o conceptos tales como clasificadores y regresiones basados en redes neuronales han emergido profusamente durante los últimos años en el vocabulario de muchas empresas como una forma de reflejar el potencial que ellas podrían alcanzar con esta tecnología aplicada a sus datos.
En este contexto, se implementó una aplicación de Minería de Datos que detecta anomalías con el fin de detectar posibles abusos en el uso de los principales servicios que otorga la empresa Aguas Araucanía S.A. Los datos analizados corresponden a las facturaciones de servicios de agua potable y alcantarillado de clientes en la ciudad de Lautaro, que fueron almacenados previamente en un Almacén de Datos. Para el análisis se utilizaron algoritmos de Minería de Datos basados en técnicas de clustering, y la metodología CRISP-DM. Como principal resultado, el sistema permitió a la empresa reducir, en forma considerable, el tiempo de búsqueda de los posibles fraudes.

1. Introducción

Si se consideran los diversos avances en las ciencias y los grandes volúmenes de datos que se han generado sólo en los últimos años, es posible notar que estos datos han sobrepasado claramente nuestra capacidad para recolectar, almacenar y comprender los mismos sin el uso de las herramientas adecuadas [1]. En este contexto es que la Minería de Datos (MD) es una de las soluciones que nos ayuda a extraer conocimiento a partir de los datos. Este conocimiento puede obtenerse a partir de la búsqueda de conceptos, ideas o patrones estadísticamente confiables, que no son evidentes a primera vista, desconocidos anteriormente y que pueden derivarse de los datos originales [2].

Dentro de las aplicaciones que se le puede dar a la MD se tienen algunas tales como análisis de fidelización de clientes, segmentación de mercados, cross-selling, optimización de la cadena de suministro, detección y prevención de fraudes [2], [3], [4], [5] y [6], detección de intrusiones en sistemas computacionales [1] y situaciones en las que se quiera analizar ciertos datos cuyo comportamiento parecen distintos del resto o también conocido como la Detección de Anomalías (DA), entre otras.

El problema de la detección de fraude, radica en el análisis de perfiles de usuario que permitan describir el comportamiento de un cliente con el fin de detectar anomalías. Es por ello que muchos de los softwares CRM (Customer Resource Management) incluyen algoritmos de MD con ese fin [7].

Es precisamente la DA, la que puede convertirse en una herramienta de enorme utilidad, que en conjunto con técnicas de Clustering, posibilitan el reconocimiento de grupos de datos cuyo comportamiento sea muy diferente al resto de los datos y también cuando no conocemos o no podemos etiquetar de manera confiable los datos para su clasificación [8].

El caso de estudio fue desarrollado en la empresa Aguas Araucanía S.A., que por tratarse de una empresa del sector de Servicios Sanitarios (agua potable, alcantarillado, tratamiento de aguas, etc.), la búsqueda de actitudes fraudulentas de los consumidores no forma parte de sus funciones activas [9].

Los administrativos, presentaron al equipo del proyecto los problemas detectados por la oficina de fraudes y se propuso utilizar técnicas de MD para la búsqueda de posibles fraudes ocasionados por clientes de la empresa.

Actualmente, la oficina de fraudes dispone de personal que revisa físicamente los medidores de los clientes caso a caso, con el fin de detectar posibles fraudes. Según declaran los administrativos de dicha oficina, se ha estimado bajo sucesos históricos, que alrededor de un 7% de los consumidores cometerían algún tipo de fraude en contra de la organización.

La empresa al no contar con herramientas que busquen activamente dichos fraudes, se encuentra en una posición en la que no puede hacer más que esperar a que el fraude se convierta en algo obvio y claramente visible, como es el caso de medidores adulterados que son detectados durante la inspección visual para el cálculo de consumo, y sólo entonces se pueden tomar medidas al respecto.

Con estos antecedentes, Aguas Araucanía S.A. ha destacado como requerimientos para el proyecto los siguientes puntos:

- Debe ser capaz de analizar y entregar indicadores sobre los siguientes datos: Consumo de servicios entregados *(agua potable, alcantarillado, etc.*), Tiempo *(año, mes, semestre, etc.)*, Ubicación *(localidad, sector, ruta)* donde se entrega el servicio y las Características del servicio

- Debe proveer capacidad de análisis visual, matemático, y entrega de reportes.

En este contexto el objetivo del artículo es presentar un estudio a través de técnicas de MD que permitan localizar y estudiar comportamientos anómalos sobre conjuntos de datos, para poder así identificar posibles fraudes en clientes que hacen uso indebido de los servicios que ofrece la empresa Aguas Araucanía S.A. Es por ello que el centro de análisis se realizó sobre el Subsistema de Facturación que pertenece al área de Gerencia de Cliente.

La estructura del artículo presenta en la segunda sección una revisión resumida de las técnicas relacionadas con Clustering y DA, luego, en la sección 3 revisamos trabajos relacionados con la DA. En la sección 4 se establece la metodología utilizada y en la sección 5 se discute sobre la experiencia revisada. Finalmente en la sección 6 se muestran las principales conclusiones, impactos y consideraciones sobre implementación de sistemas para DA y su aplicación en el caso de estudio.

2. Clustering y DA

La meta principal en la DA, es encontrar objetos que sean diferentes de los demás. Frecuentemente estos objetos son conocidos como *Outlier* [10] [11]. La DA también es conocida como detección de desviaciones, porque objetos anómalos tienen valores de atributos con una desviación significativa respecto a los valores típicos esperados. Aunque estas Anomalías son frecuentemente tratados como ruido o error en muchas operaciones, tales como Clustering, por ejemplo para propósitos de detección de fraude son una herramienta valiosa en la búsqueda de comportamientos atípicos [12], [13], [14].

En términos de cómo identificar y agrupar estos datos anómalos, pueden clasificarse según [10] en técnicas basadas en:

- **Modelos:** Se basan en el campo de la estadística, dada la premisa de conocer la distribución de los datos.
- **Proximidad:** Se basan fundamentalmente en el manejo de distancias entre objetos, entre mayor sea la distancia del objeto respecto a los demás, éste es considerado como una Anomalía.
- **Densidad: S**e basan en la estimación de densidad de los objetos, para ello, los objetos localizados en regiones de baja densidad y que son relativamente distantes de sus vecinos se consideran anómalos. La principal característica de ésta es que generalmente son de aprendizaje no supervisado, pues en la mayoría de los casos, no se conoce la clase, para ello se asigna una calificación a cada instancia que refleja el grado con el cual la instancia es anómala.

Una de las herramientas necesarias para la DA es el Clustering, que consiste en agrupar un conjunto de datos, sin tener clases predefinidas, basándose en la similitud de los valores de los atributos de los distintos datos. Esta agrupación, a diferencia de la clasificación, se realiza de forma no supervisada, ya que no se conoce de antemano las clases del conjunto de datos de entrenamiento. [15], [16], [17].

El Clustering identifica clusters, o regiones densamente pobladas, de acuerdo a alguna medida de distancia, en un gran conjunto de datos multidimensional [18]. El Clustering se basa en maximizar la similitud de las instancias en cada cluster y minimizar la similitud entre clusters [19].

Dentro del análisis de Clustering existen, básicamente, los siguientes tipos de métodos: los jerárquicos, los de partición, los basados en densidad, los métodos basados en cuadrículas, los basados en restricciones y los escalabres [20]. En el caso de los primeros, se intenta ordenar los elementos a distintos niveles de similitud; mientras que los segundos, meramente, asignan cada elemento a un grupo de manera tal de obtener conjuntos

homogéneos [21]. De los métodos jerárquicos, podemos agregar que estos métodos son sensibles a los elementos outliers. Consecuentemente, la configuración final de los grupos debe ser cuidadosamente analizada, para detectar tales situaciones.

K-Means [22] en un método particional donde se construye una partición de una base de datos D de n objetos en un conjunto de k grupos, buscando optimizar el criterio de particionamiento elegido. En K-Means cada grupo está representado por su centro. K-Means intenta formar k grupos, con k predeterminado antes del inicio del proceso. Asume que los atributos de los objetos forman un vector espacial. El objetivo que se intenta alcanzar es minimizar la varianza total intra-grupo o la función de error cuadrático [23]. Una desventaja importante de este método es su alta dependencia de la partición inicial o la selección inicial de los puntos-centro. Si dos o más puntos-centro caen dentro de lo que conformaría un mismo cluster, ambos grupos estarían pobremente diferenciados. Asimismo la existencia de outliers, produce al menos un grupo con elementos demasiado dispersos. Además, el no conocer el valor de K puede dar lugar a agrupamientos no naturales. Por estas razones, para mejorar la estabilidad del método, es deseable volver a correr el algoritmo con otras configuraciones iniciales [21].

COBWEB [21, 24] se trata de un algoritmo jerárquico, que se caracteriza porque utiliza aprendizaje incremental, esto es, realiza las agrupaciones instancia a instancia. Durante la ejecución del algoritmo se forma un árbol (árbol de clasificación) donde las hojas representan los segmentos y el nodo raíz engloba por completo el conjunto de datos de entrada. Al principio, el árbol consiste en un único nodo raíz. Las instancias se van añadiendo una a una y el árbol se va actualizando en cada paso. La actualización consiste en encontrar el mejor sitio donde incluir la nueva instancia, operación que puede necesitar de la reestructuración de todo el árbol (incluyendo la generación de un nuevo nodo anfitrión para la instancia y/o la fusión/partición de nodos existentes) o simplemente la inclusión de la instancia en un nodo que ya existía. La clave para saber cómo y dónde se debe actualizar el árbol la proporciona una medida denominada utilidad de categoría, que mide la calidad general de una partición de instancias en un segmento. La reestructuración que mayor utilidad de categoría proporcione es la que se adopta en ese paso. El algoritmo es muy sensible a otros dos parámetros: *acuity* que representa la medida de error de un nodo, y c*ut-off* que es un valor que se utiliza para evitar el crecimiento desmesurado del número de segmentos indicando el grado de mejoría que se debe producir en la utilidad de categoría para que la instancia sea tenida en cuenta de manera individual.

Expectation-Maximization (EM) [21] pertenece a una familia de modelos que se conocen como *Finite Mixture Model*, los cuales se pueden utilizar para segmentar conjuntos de datos. Es un método de clustering probabilístico. Se trata de obtener la Función de Densidad de Probabilidad (FDP) desconocida a la que pertenecen el conjunto completo de datos. Esta FDP se puede aproximar mediante una combinación lineal de NC componentes, definidas a falta de una serie de parámetros que son los que hay que averiguar. Cada cluster se corresponde con las respectivas muestras de datos que pertenecen a cada una de las densidades que se mezclan. Se pueden estimar FDP de formas arbitrarias, utilizándose FDP normales n-dimensionales, t-Student, Bernoulli, Poisson, y log-normales. Aquí se modelarán los datos mediante distribuciones normales, por ser éstas las más comunes.

Otros algoritmos usados en el análisis de Clúster se encuentra: DBSCAN, Sequence, Kohonen, TwoStep [25-27].

Debilidades propias del análisis basados en clúster tienen que ver con la posibilidad de hacer una mala elección de métricas, tal como lo propone [4, 5]. En este sentido, puede ser dificultoso combinar variables continuas y aquellas que representen categorías.

3. Trabajos relacionados

El uso de la MD para la detección de anomalías con el fin de detectar fraudes puede ser muy variado, encontrándose distintos tipos de aplicaciones en la literatura actual. Es así como en [2] se realiza una investigación exhaustiva de uso de la MD para la detección de fraudes, definiendo los tipos y subtipos, métodos y técnicas para la detección de fraudes, así como las limitaciones de éstos.

Algo similar ocurre en [6] que propone una aplicación de la MD para la detección de fraudes en subastas por Internet, usando para ello análisis de redes sociales y árboles de decisión. La idea es analizar patrones de relaciones e interacción entre participantes de la red, con el fin de descubrir estructuras sociales subyacentes.

Por otro lado, [4] propone el análisis de grupo de pares para monitorear el comportamiento en el tiempo en el uso de tarjetas de crédito con el fin de buscar posibles fraudes.

Una investigación diferente realiza [3] al definir un modelo de costos que permite maximizar los beneficios versus minimizar los costos que significa realizar una auditoría en el ámbito fiscal. Propone el uso de árboles de decisión para la búsqueda de anomalías.

Un trabajo interesante es el realizado por [28] quien propone una nueva definición de anomalías llamada Cluster de Anomalías, la cual se basa en el hecho de que muchos eventos que podrían considerarse anormales para un conjunto de datos pueden agruparse en pequeños cluster de anomalías.

El autor [1] hace un recuento de aplicaciones de MD usadas para la detección de intrusos en sistemas informáticos. Dentro de este recuento se puede destacar el concepto de Minería de Anomalías, que consiste en hacer un análisis de éstos datos anómalos dentro de un conjunto.

En la literatura, no se encontraron ejemplos de detección de anomalías para el uso indebido de servicios sanitarios. Favorablemente, el análisis que se debe realizar es similar a la detección de fraudes en tarjetas de crédito al disponer de un conjunto de clientes que presenta comportamientos distintos al normalmente establecido por las empresas del rubro. A continuación se describen los materiales y métodos utilizados para la detección de anomalías en el uso de servicios ofrecidos por la empresa Aguas Araucanía S.A.

4. Materiales y Métodos utilizados

Para el desarrollo del trabajo se utilizó Clementine Client[1], herramienta líder y conocida mundialmente, que posee potentes herramientas de visualización y una gran variedad de técnicas de aprendizaje automático para clasificación, regresión, Clustering y discretización entre otras, entregando apoyo completo para el ciclo de MD a través de la metodología CRISP-DM (CRoss-Industry Standard Process for Data Mining) [29], lo que reduce el tiempo de entrega de la solución final.

CRISP-DM es una metodología que ha sido desarrollada para la construcción de proyectos de MD. Propone un ciclo de vida que consiste en seis etapas. La secuencia de los mismos no es rígida, siempre se requiere desplazarse hacia delante y hacia atrás entre las etapas [29]. Estas etapas serán utilizadas para el desarrollo del proyecto.

Es destacable que Clementine contiene múltiples algoritmos para la detección de fraudes, entre los cuales se encuentran los de Clusterización y de Detección de Anomalías (Outlier). El software analiza los resultados obtenidos con el set de datos introducidos y busca la mejor alternativa con el menor error posible al aplicar cada uno de los algoritmos.

El desarrollo del proyecto contempló las etapas sugeridas por la metodología CRISP-DM, las que se detallan a continuación:

4.1 Comprensión del negocio

Para comprender el negocio, fue necesario realizar una serie de reuniones con la Gerencia de Clientes. Los administrativos, presentaron al equipo del proyecto los problemas detectados por la oficina de fraude, los que fueron estudiados con el fin de buscar alternativas de solución. El equipo propone utilizar técnicas de Clustering para la detección de anomalías con el fin de obtener un listado de clientes que presenten datos atípicos.

Actualmente, la oficina de fraudes dispone de personal que revisa físicamente los medidores de los clientes caso a caso, con el fin de detectar posibles anomalías. Para ello, en cada inspección se selecciona una determinada ruta (subsector), la que es revisada por completo. Este proceso es lento y no asegura la detección oportuna de las posibles anomalías, dado que no siempre es fácil detectarlos.

4.2 Comprensión de los datos

En esta etapa fue necesario analizar el modelo de base de datos relacional del sistema de información de la empresa, específicamente aquellas entidades que tienen relación con el proceso de facturación. Éste mantiene información del consumo mensual de uso de agua potable y alcantarillado de cada cliente que utiliza estos servicios. Obtener el conjunto de datos a analizar no es un proceso trivial, por tanto fue necesario reunirse con personal especializado de la empresa, revisar documentación de la base de datos, revisar nombres de atributos y el diccionario de datos; entre otros.

4.3 Preparación de los datos

En cuanto a los datos de análisis, debió diseñarse un Almacén de Datos (AD) que se alimente de las bases de datos transaccionales a través de un proceso de extracción de datos previamente definido llamado ETL. Los datos son almacenados en un repositorio que consiste en hechos y dimensiones representados a través de un esquema en estrella. La tabla de hechos almacena los indicadores a medir y las dimensiones representan los criterios de análisis. Cuando se mantiene una estructura de un AD, pero adaptada sólo a un sector de la empresa, o para un fin concreto, se utiliza un Data Mart que es parte del AD completo [18].

En la Figura 1 se presenta el Data Mart creado para el área de estudio, Gerencia de Clientes.

En cuanto al Data Mart presentado, éste contiene información específica sobre los consumos históricos de agua potable y alcantarillado de la región, ya que son datos que dispone la empresa. Las variables que lo

[1] software que pertenece a SPSS http://www.spss.com.

componen son: la tabla de hechos, que almacena los datos de unidades (m3) facturadas de consumo de agua potable, alcantarillado, el sobreconsumo de agua potable y un contador de consumos facturados; por otro lado, las tablas de dimensiones que contienen información de las distintas unidades de tiempo, lugares y características del servicio.

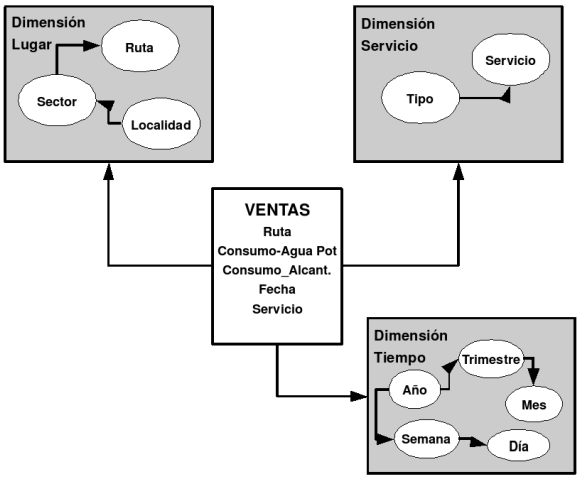

Fig. 1 Data Mart para Gerencia de Clientes

Los datos almacenados presentan la siguiente información:

- **Lugar**: que indica la localidad con su sector y la ruta (sector pequeño).

- **Servicio**: que detalla el tipo de servicio prestado

- **Tiempo**: esta dimensión es necesaria para analizar lo que ha ocurrido históricamente y así proyectar al futuro.

- **Tabla de hechos Venta**: Que almacena los indicado-res a consultar.

En conjunto con personal técnico de la empresa fue posible adquirir los datos históricos de los consumos facturados desde el inicio del funcionamiento de dicho sistema de facturación hasta el mes de octubre del año 2007. Los datos fueron cargados en el Data Mart a través de un proceso ETL diseñado para ello.

Para el caso de estudio se utilizó el 4% de los datos almacenados en el Data Mart, los que pertenecen a las facturaciones emitidas en la localidad de Lautaro. Como conjunto de entrenamiento o training set, se utilizó el 50% de los datos de dicha localidad, y el resto de los mismos, fueron utilizados para hacer los test.

4.4 Modelado

El modelo de detección de anomalías de Clementine, entrega como resultado grupos de datos con características similares, los cuales son llamados grupos homólogos del modelo. Cada grupo homólogo entrega información sobre la cantidad de registros procesados, la cantidad de anomalías encontradas, un resumen sobre los campos escogidos a estudiar, entre otros.

Cabe destacar que la cantidad generada de grupos homólogos va a depender directamente de los parámetros de configuración del modelo, pues con modificar un solo parámetro, no se generará la misma cantidad de grupos y por lo tanto los resultados serán distintos. Ejemplo de parámetros a introducir son: agregar dimensiones al AD, índice de anomalías, cantidad de grupos que se desee obtener, selección de algoritmos de análisis, etc.

En la Figura 2, Anomaly1 contiene la configuración de los parámetros necesarios para que el software genere los grupos homólogos. Toda la información que se genera al momento de ejecutar los algoritmos serán almacenados en un archivo (Tabla) para su análisis posterior.

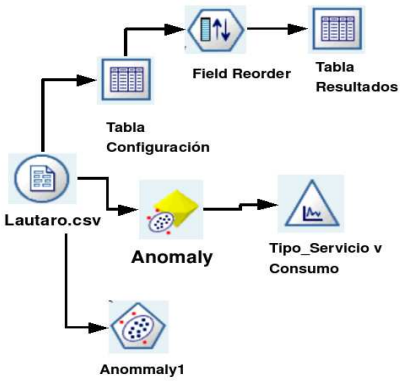

Fig. 2. Modelo de DA diseñado para Gerencia de Clientes.

Los algoritmos que se seleccionaron para realizar este análisis son: K-means, COBWEB y EM.

Una vez aplicado el modelo al conjunto de datos, Clementine genera los grupos homólogos. Un ejemplo es el representado en la Figura 3, en el que se ha utilizado el algoritmo K-means con un índice de anomalía de 0.5. Es posible destacar los casos normales en azul (valor F) y los anómalos en rojo (valor T). Los casos en color rojo representan registros de clientes que podrían presentar anomalías frente a su grupo.

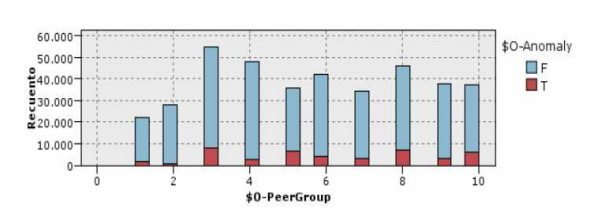

Fig. 3. Grupos Homólogos entregados por Clementine.

4.5 Evaluación

A través de los algoritmos de detección de anomalías de Clementine Client se probaron los datos almacenados en el Data Mart, entregando una serie de resultados que deben ser analizados cada vez que se inicia el ciclo de la metodología CRISPDM.

Como se mencionó anteriormente, COBWEB admite dos parámetros, *cutoff* y *acuity*. Los valores que Clementine propone para éstos es de *acuity*=1.0 y *cutoff*=0.0028. Con estos valores no se obtiene ningún cluster, por lo que es necesario modificarlos. Parece lógico decrementar *cutoff* para obtener un mayor número de clusters, para ello se le asigna un valor de *cutoff*=0.00028, obteniéndose 1292 clusters. Cada cluster está formado por unos pocos elementos, de manera que se tienen muchísimos clusters con pocos registros. El número tan elevado de clusters hace evidente que el valor elegido no es el más adecuado, es necesario incrementarlo. Tras varios intentos se utiliza un valor de *cutoff*=0.0018, obteniéndose 74 clusters. Se hicieron nuevas pruebas, pero ninguna ofreció un número de clusters menor. De los clusters obtenidos, 73 cubren 856.324 registros repartidos uniformemente, mientras que el cluster restante engloba al resto de los registros.

EM y K-means proporcionaron una segmentación con 10 clusters, muy similar al de la figura 3, por lo que se consideran efectivos.

Los registros anómalos encontrados deben ser contrastados con la información histórica de casos de fraude que almacena la empresa, con el fin de verificar que la información entregada por el software sea la correcta. En cada iteración, es necesario introducir nuevos parámetros o modificar algunos ya existentes para realizar un nuevo análisis. Es así que para K-means se utilizó un índice de anomalías de 0.5, 1 y 2. Para el caso de EM se configura la cantidad máxima de iteraciones, que es 100.

4.6 Implementación

Finalmente se implementa el sistema de detección de anomalías en el servidor de la empresa, con el fin de obtener listados de clientes que podrían ser casos de análisis. Este listado es utilizado por la oficina de fraudes de Aguas Araucanía S.A. para corroborar en terreno cada uno de los casos. Con ello, la oficina de fraude sólo debe preocuparse de revisar los casos expuestos en la lista, y no de todas las unidades que componen una ruta o sector.

5. Resultados y discusión

Dentro de los resultados obtenidos, se debe destacar que los casos detectados como anómalos no necesariamente se tratan de casos de fraudes, ya que es posible que falte agregar nuevos parámetros al software.

COBWEB ofrece un único cluster significativo, con resultados no muy diferentes de los que se obtendrían sin realizar segmentación alguna, considerando todos los registros para una única curva de regresión. Este algoritmo no es adecuado para la detección de registros anómalos, ya que tiende a agrupar a la mayoría de los registros en un solo segmento.

EM y k-medias ofrecen segmentos de similares formas, aunque han agrupado los registros de diferente forma. Ello se debe al hecho de que ambos forman parte de una misma familia de algoritmos de Clustering y tienen bases comunes. Sin embargo EM destaca claramente sobre k-medias, ofreciendo mejores resultados al momento de validar los registros encontrados como anómalos.

Al utilizar por primera vez EM se tuvo que un 40% de los registros presentados como anómalos son incorrectos, no correspondiendo a casos de fraude, por lo que fue necesario introducir una mayor cantidad de registros para realizar un nuevo análisis. Después de varios intentos, se obtuvo información confiable. Cerca del 73% de los registros detectados como anómalos son posibles causas de fraude. Esto permitirá a la empresa verificar rápidamente el posible fraude revisando los medidores de clientes.

Para K-means se obtuvieron resultados similares, sólo que el porcentaje de registros que posiblemente sean casos de fraude no supera el 55%.

EM y k-medias, perteneciendo a la familia de algoritmos de particionado y recolocación, ofrecen mejores resultados que COBWEB, indicando que para la detección de anomalías son más adecuados que éste.

Cabe destacar que para el futuro, será necesario introducir datos que reflejen la estación del año (Primavera, Verano, Otoño e Invierno) y la temperatura. Está comprobado que en época de verano los usuarios

del servicio consumen una mayor cantidad de agua, cosa que actualmente no es detectada por el modelo.

El sistema de DA diseñado para el caso de estudio permite a la empresa reducir, en forma considerable, el tiempo de búsqueda de los posibles fraudes. También permitió reducir los costos de administración de la oficina de fraudes, ya que el proceso de detección de anomalías ha sido mejorado.

6. Conclusiones

La MD es una técnica eficiente para la detección de anomalías, siempre y cuando se disponga de un conjunto de datos suficientes para un correcto análisis y una metodología que permita llevar un control de los resultados dando la posibilidad de reestructurar medidas como la: recolección de nuevos datos, separación de datos en clases, transformaciones de las variables, eliminación de datos, selección de otros algoritmos de MD, cambio en los parámetros introducidos en los modelos, delimitación del campo de búsqueda, etc.

Según [7], la MD aporta diferentes tecnologías en la identificación de operaciones fraudulentas. Por lo general es necesario el uso de varias de estas tecnologías, con el fin tener un mejor éxito en la solución del problema. La elección exacta y la combinación de estas tecnologías, depende en gran medida de las características de los datos disponibles. Para el caso de estudio se verifica que la herramienta Clementine es apropiada para la detección de fraudes dado que dispone de algoritmos eficientes de Clusterización y detección de Anomalías, utilizando para ello la metodología CRISP-DM para el diseño de modelos de MD.

Se puede concluir que el algoritmo EM, siendo un algoritmo que realiza C*lustering* probabilístico, es más adecuado que el algoritmo k-medias y COBWEB, para segmentar los datos del AD diseñado para la empresa Aguas Araucanía S.A., con el fin de encontrar posibles casos de fraude.

El sistema creado permite a la empresa disponer de una lista de clientes que presentan comportamientos anómalos, dando la posibilidad de detectar posibles fraudes en forma oportuna. En este sentido, se obtuvo una reducción del tiempo de búsqueda y del costo asociado para ello.

7. Referencias

[1] A. Singhal and S. Jajodia, " Data Modeling and Data Warehousing Techniques to Improve Intrusion Detection," *Serie Libros Advances in Information Security*, vol. 31, pp. 69-82, 2007.

[2] C. Phua, V. Lee, K. Smith, and R. Gayler, "A Comprehensive Survey of Data Mining-based Fraud Detection Research," *Clayton School of Information Technology, Monash University*, 2005.

[3] F. Bonchi, F. Giannotti, G. Mainetto, and D. Pedreschi, "Using Data Mining Techniques in Fiscal Fraud Detection," *LNCS 1676*, pp. 369-376, 1999.

[4] R. Bolton and D. Hand, "Unsupervised Profiling Methods for Fraud Detection," *Credit Scoring and Credit Control VII*, 2001.

[5] R. Bolton and D. Hand, "Statical Fraud Detection: A Review (With Discussion)," *Statistical Science*, vol. 17, pp. 235-255, 2002.

[6] Y. Ku, Y. Chen, and C. Chiu, "A Proposed Data Mining Approach for Internet Auction Fraud Detection," *LNCS 4430*, pp. 238-243, 2007.

[7] W. Santamaría, "Técnicas de minería de datos para la detección de fraude," *Maestría en Ingeniería de Sistemas y Computación- Universidad Nacional de Colombia.*, 2008.

[8] K. Niu, C. Huang, S. Zhang, and J. Chen, "ODDC: Outlier Detection Using Distance Distribution Clustering," *LNAI*, vol. 4819, pp. 332-343, 2007.

[9] A. Cravero and S. Sepúlveda, "Detección de fraudes con el uso de Minería de Datos, un caso de estudio en empresa de Servicios Sanitarios.," presented at VIII Congreso Chileno de Investigación Operativa, OPTIMA 2009, 2009.

[10] L. Davies and U. Gather, "The Identification of Multiple Outliers.," *Journal of the American Statistical Association*, vol. 88, pp. 782-792, 1993.

[11] C. Caroni and P. Prescott, "On Rohlf's Method for the Detection of Outliers in Multivariate Data," *Journal of Multivariate Analysis*, vol. 52, pp. 295-307, 1995.

[12] J.-X. Pan, W.-K. Fung, and K.-T. Fang, "Multiple outlier detection in multivariate data using projection pursuit techniques," *Journal of Statistical Planning and Inference*, vol. 83, pp. 153-167, 2000.

[13] K. Das and J. Schneider, "Detecting anomalous records in categorical datasets," presented at Proceedings of the 13th ACM SIGKDD International, 2007.

[14] T. Hu and S. Sung, "Detecting pattern-based outliers," *Pattern Recognition Letters*, vol. 24, pp. 3059-3068, 2003.

[15] Z. He, X. Xu, and S. Deng, "A Fast Greedy Algorithm for Outlier Mining," *Computer Science*, 2005.

[16] A. C. Atkinson and M. Riani, "Exploratory tools for clustering multivariate data," *Computational Statistics and Data Analysis.*, vol. 52, pp. 272-285, 2007.

[17] N. Wu and J. Zhang, "Factor-analysis based anomaly detection and clustering," *Decision Support Systems*, vol. 42, pp. 375-389, 2006.
[18] M. Chen and J. Han, "Data mining: An overview from database perspective.," *IEEE Transactions on Knowledge and Data Eng*, 1996.
[19] J. Han and M. Kamber, "Data mining: Concepts and techniques," *Morgan Kauffmann Publishers*, 2001.
[20] M. Garre and M. Charro, "Estimación del esfuerzo de un proyecto Software utilizando el criterio MDL-EM y Componentes normales n-dimensionales. Aplicación a un caso práctico.," *Revista de Procesos y Métricas de las Tecnologías de la Información (RPM)*, vol. 2, pp. 13-24, 2005.
[21] J. A. Soto, I. Ponzoni, and G. Vazque, "Análisis Numérico de diferentes criterios de similitud en Algoritmos de Clustering," *Mecánica Computacional*, vol. 25, pp. 993-1011, 2006.
[22] P. Britos, A. Hossian, R. García-Martinez, and E. Sierra, "Minería de Datos Basada en Sistemas Inteligentes," 2005.
[23] F. Valenga, E. Fernández, H. Merlino, D. Rodríguez, C. Procopio, P. Britos, and R. García-Martínez, "Minería de Datos Aplicada a la Detección de Patrones Delictivos en Argentina," *VII Jornadas Iberoamericanas de Ingeniería del Software e Ingeniería del Conocimiento*, 2008.
[24] D. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Machine Learning*, vol. 2, pp. 139-172, 1987.
[25] T. Chiu, D. Fang, J. Chen, Y. Wang, and C. Jeris, "A robust and scalable clustering algorithm for mixed type attributes in large database environment," presented at Proceedings of the seventh ACM SIGMOD International conference on Knowledge Discovery and Data Mining., 2001.
[26] M. T. S. De Lejarza, "Redes neuronales auto-organizadas y clustering: Una aplicación a la agrupación económico-funcional de entidades de población," *Quaderns de Treball*, 1996.
[27] C. Chiu and C. Tsai, "A web Services-Based Collaborative Scheme for Credit Card Fraud Detection.," *Proc. of 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service*, 2004.
[28] L. Duan, L. Xu, Y. Liu, and J. Lee, "Cluster-based outlier detection," *Springer Science + Business Media, LLC*, 2008.
[29] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0 step-by-step data mining guide," *CRISPDM*, 2000.