International Journal of Information Systems and Software Engineering for Big Companies (IJISEBC)

Recibido: 03-10-2019 / Revisado: 23-11-2019
Aceptado: 23-11-2019 / Publicado: 15-12-2019

Modelado y simulación de robos y hurtos basados en redes SOM, TDIDT y Bayesianas. Un caso de estudio

Modeling and simulation of robberies and thefts based on SOM, TDIDT and Bayesian networks. A case study

Lorena E. Flores[1], Sonia I. Mariño[1], Sebastian Martins[2]

[1] Universidad Nacional del Nordeste, Argentina
[2] Universidad Nacional de Lanús, Argentina

lorenaelizabeth.flores@gmail.com , simarinio@yahoo.com , smartins089@gmail.com

RESUMEN. Se presenta la integración de tecnologías de minería de datos y GIS, orientadas a la generación de conocimiento para identificar y caracterizar clusters de robos y hurtos en una ciudad argentina en el primer semestre de 2017. Se adaptó la metodología CRISP-DM, y se aplicó un conjunto de técnicas de minería de datos (SOM, TDIDT y Redes Bayesianas) para identificar y comprender los patrones delictivos. Además, se vincularon los patrones descubiertos con la tecnología GIS para comprender las zonas calientes de mayor ocurrencia de estos delitos. La finalidad es proponer innovadoras modalidades para apoyar procesos de decisión basados en TI.

ABSTRACT. The integration of data mining and GIS technologies is presented, in order to generate knowledge to identify and characterize theft and robbery clusters in an Argentine city in the first half of 2017. The CRISP-DM methodology was adapted, and applied to a set of data mining techniques (SOM, TDIDT and Bayesian Networks) to identify and understand criminal patterns. In addition, the patterns discovered were linked with GIS technology to understand the hot zones with the highest occurrence of these crimes. The purpose of the paper is present innovative modalities to support IT-based decision processes.

PALABRAS CLAVE: Minería de datos, GIS, Toma de decisiones, Aplicaciones en gobierno, Redes SOM, TDIDT y Bayesianas.

KEYWORDS: Data mining, GIS, Decision making, Government applications, SOM, TDIDT and Bayesian Networks.

Flores, L. E.; Mariño, S. I.; Martins, S. (2019). Modelado y simulación de robos y hurtos basados en redes SOM, TDIDT y Bayesianas. Un caso de estudio. *International Journal of Information Systems and Software Engineering for Big Companies (IJISEBC)*, 6(2), 81-87.

1. Introducción

La inteligencia de negocios o Business Intelligence (BI, por sus siglas en inglés) ofrece un concepto clave desde una perspectiva empresarial, combina tecnologías y procesos que asisten al análisis de los datos y a la transformación de información empresarial, con el fin de obtener conocimiento válido que permita la toma de decisiones apropiadas y orientadas a optimizar los recursos y mejorar los resultados.

La minería de datos (MD) es la subdisciplina de los sistemas de información que contribuye a la inteligencia de negocios las herramientas necesarias para explotar la información necesaria y transformarla en conocimiento útil. Para realizar este proceso de búsqueda de patrones se requiere el uso de técnicas de minería de datos (agrupación, árboles de decisión, redes neuronales artificiales, etc.) y la aplicación de algoritmos específicos dependiendo de la problemática (agrupamiento, clasificación, predicción, entre otros) (Anoopkumar & Rahman, 2016).

Existen diversos procesos de minería o explotación de información para obtener conocimiento a partir de los datos disponibles (García-Martínez, Britos & Rodríguez, 2013):

- Descubrimiento de reglas de comportamiento: El proceso de descubrimiento de reglas de comportamiento aplica cuando se requiere identificar cuáles son las condiciones para obtener determinado resultado en el dominio del problema.
- Descubrimiento de grupos: Este proceso se aplica cuando se requiere identificar una partición en la masa de información disponible sobre el dominio de problema.
- Ponderación de interdependencia de atributos: Este proceso se aplica cuando se requiere identificar cuáles son los factores con mayor incidencia (o frecuencia de ocurrencia) sobre un determinado resultado del problema.
- Descubrimiento de reglas de pertenencia a grupos: Este proceso se aplica cuando se requiere identificar cuáles son las condiciones de pertenencia a cada una de las clases en una partición desconocida "a priori", pero presente en la masa de información disponible sobre el dominio de problema.
- Ponderación de reglas de comportamiento o de la pertenencia a grupos: Este proceso se aplica cuando se requiere identificar cuáles son las condiciones con mayor incidencia (o frecuencia de ocurrencia) sobre la obtención de un determinado resultado en el dominio del problema, sean estas las que en mayor medida inciden sobre un comportamiento o las que mejor definen la pertenencia a un grupo.

En el análisis de delitos criminales, el descubrimiento de patrones significativos ha brindado la posibilidad de obtener datos de interés para interpretar y adecuar el conocimiento en la definición de los planes de prevención requeridos. La aplicación de diversas técnicas de minería de datos sobre el campo criminal se ha convertido en una herramienta con un gran potencial que permite diseñar estrategias específicas para esta área, resultando en un proceso automático de extracción de conocimiento útil (McCue, 2014).

Así mismo, el desarrollo de nuevas tecnologías han logrado implementar nuevas herramientas como el GIS (Geographic Information System o Sistema de Información Geográfica, SIG su acrónimo en castellano), tal como lo define Chang (Chapman et al., 2000), "…un sistema informático que permite capturar, almacenar, consultar, analizar y mostrar datos geoespaciales". Desde su lanzamiento, el GIS ha logrado traspasar las barreras tecnológicas presentando la información espacial a través de mapas georreferenciados. Estos mapas juegan un papel esencial en la vida cotidiana de las personas, su incorporación permite analizar, comunicar y compartir información con el fin de resolver y abordar problemáticas diarias, y ayudan a asistir a la toma de decisiones más inteligentes.

Desde esta perspectiva, el GIS puede tratarse como una base de datos que contiene información geográfica valiosa vinculada a un determinado territorio de interés (Rakotomalala, 2004). Por ello, siguiendo a Baghdadi, Mallet y Zribi (2018), resulta viable poder aplicar MD sobre base de datos georreferenciadas, para el hallazgo de patrones y regularidades significativas que resulten en conocimiento de interés sobre el territorio asociado.

En el análisis delictivo, la MD aplicada sobre datos espaciales criminales permite la visualización de los hechos delictivos a través de mapas, con el fin de analizar, clasificar y predecir tendencias del delito de acuerdo a una determinada zona geográfica.

Este trabajo se validó en la detección y el análisis del comportamiento de hechos delictivos de robos y hurtos que afectan en una ciudad en el período comprendido por el primer semestre del año 2017, aplicando técnicas de minería de datos sobre una base de datos georreferenciada, y apoyados en tecnología GIS e Infraestructura de Datos Espaciales (IDE) para visualizar la información.

El conocimiento obtenido de la integración de la tecnología GIS y algunos algoritmos de la minería de datos se reflejó en la construcción del denominado Mapa del Delito o Mapa del Crimen de la ciudad. Ésta es una herramienta cartográfica que permitió mapear y visualizar los patrones de delictualidad obtenidos del análisis realizado, en el cual se identificaron las zonas de mayor riesgo de ocurrencia de hechos según distintos criterios.

Así, la implementación de esta solución tecnológica permitió identificar patrones para detectar y predecir la ocurrencia de estos tipos de delitos, relacionados con ubicaciones geográficas y otras variables de interés, entre las que se mencionan: i) características del hecho: lugar, día y horario de ocurrencia del delito, tipo de elemento sustraído (vehículo, domiciliario, otros), tipo de ataque (forcejeo, arrebato, etc.), tipo de arma u objeto utilizado en la escena y jurisdicción policial interviniente, ii) registro del autor del hecho: características del sospechoso (sexo del delincuente, edad aproximada, etc.), iii) registro del denunciante: sexo y edad de la víctima, entre otras.

Mediante este análisis es posible obtener información para la toma de decisiones, orientada el diseño de diferentes planes de prevención, mejora de la seguridad, alerta de situaciones de riesgo al ciudadano, reducción del impacto de la delincuencia, y mayor eficacia y eficiencia en el accionar de las fuerzas policiales ante estos actos delictivos, entre otros posibles impactos positivos.

2. Metodología

En la construcción de la solución propuesta, se contemplaron:

2.1. Adaptación de la metodología de CRISP-DM

En la literatura se mencionan diversas metodologías para modelar y procesar procesos de minería de datos. En este trabajo se optó por la metodología de CRISP-DM (Chapman et al., 2000) para explotar los datos, que consta de las siguientes fases:

- Comprensión del negocio: Se determinan los objetivos y requerimientos del proyecto desde una perspectiva del negocio, definiendo el problema de minería y el plan de trabajo.
- Comprensión de los datos: Se aborda la recolección de datos que se utilizarán en el proyecto y la familiarización con los mismos. Podrían surgir primeras hipótesis acerca de la información que podría estar oculta.
- Preparación de los datos: Comprende actividades de tratamiento de los datos para construir el conjunto final sobre el cual se aplicarán las técnicas de minería.
- Modelado: Se aplican las diversas técnicas y algoritmos de minería sobre el conjunto de datos para generar información y los patrones implícitos en ellos.
- Evaluación: Se analizan los patrones obtenidos en función de los objetivos organizacionales. Se debería determinar si se ha omitido algún objetivo importante del negocio y si el nuevo conocimiento será implementado, es decir, si se pasará a la próxima etapa.
- Implementación: Se comunica el nuevo conocimiento con fines de utilización. Éste debe ser representado de forma entendible para el usuario.

2.2. Algoritmos para modelar y explotar los datos

Se propone la aplicación de los procesos de explotación de información que comprende la combinación de las siguientes técnicas de minería de datos:

- SOM y TDIDT aplicados al descubrimiento de Reglas de Pertenencia a Grupos: Para el descubrimiento de reglas de pertenencia a grupos se propone utilizar mapas auto-organizados (SOM) y, una vez identificados los grupos, aplicar algoritmos de inducción (TDIDT) con el objeto de establecer las reglas de pertenencia a cada uno. Se optó por el algoritmo Kohonen-SOM para el descubrimiento de grupos y el algoritmo C4.5 para la caracterización o descubrimiento de reglas de cada clúster.
- Redes Bayesianas aplicadas a la Ponderación de Interdependencia entre Atributos: se opta por Redes Bayesianas Para determinar en qué medida la variación de los valores de un atributo incide sobre la variación del valor de un atributo clase. Se utilizó el algoritmo Naive Bayes para ponderar los atributos.

2.3. Software para explotar los datos

Se optó por Tanagra para ejecutar los procesos de explotación de información y generar información. Ofrece varios métodos de minería de datos de análisis exploratorio de datos, aprendizaje estadístico, aprendizaje automático y área de bases de datos. Implementa varios algoritmos de aprendizaje supervisado y no supervisado (Rakotomalala, 2004).

3. Resultados

Para la detección de los delitos de robo y hurto cometidos en una ciudad en el periodo de tiempo de Enero–Junio de 2017, se utilizó una base de datos delictiva recibida del SAT (Sistema de Alerta Temprana), un sistema informático que almacena y concentra datos criminales referentes a los distintos tipos de delitos ocurridos en Argentina, y a partir del cual se realiza un análisis estadístico de la información. Actualmente se desconoce el uso de técnicas o herramientas de minería de datos aplicadas a estos datos.

Se estableció como objetivo de minería de datos, que sustenta el modelado y experimentación expuesta en el trabajo:

- Analizar una muestra de datos correspondientes al primer semestre del año 2017 registrados para la ciudad y generar patrones de comportamiento relevantes a través de la agrupación de casos. Definidos los grupos, se deben determinar los factores que caracterizan la ocurrencia de delitos de robo y hurto en la ciudad, y detectar aquellos con mayor incidencia dentro de cada grupo.

Para lograr dicho objetivo se establecieron las siguientes necesidades:

- Identificar y caracterizar grupos que definan el comportamiento de delitos en base a las características temporales y del delito, para comprender con mayor detalle indicadores que definan a dichos grupos.
- A partir del comportamiento definido en el punto anterior, determinar cual/es de las características tiene un mayor nivel de incidencia en la ocurrencia de delitos en la ciudad.

Consideraciones:

- Los datos se obtuvieron de la base de datos gestionada/administrada por el SAT.
- La construcción del modelo, que responde al objetivo de minería de datos definido, implica: establecer los valores de los parámetros para ejecutar las técnicas "SOM y TDIDT aplicados al descubrimiento de reglas de pertenencia a grupos" y "Redes Bayesianas aplicadas a la ponderación de interdependencia entre atributos".

Se presentan los resultados del descubrimiento de reglas de pertenencia a grupos que identifican y caracterizan el comportamiento del delito. Se contempla:

Formación de clusters

Para la formación de grupos se definen los atributos de entrada requeridos por el algoritmo Kohonen SOM: dia_m, mes_r, horas_r, delito_descrip, tipo_lugar, clase_arma, elemento_sustraido, tipo_ataque. Se aplicó el algoritmo y se obtuvieron cuatro grupos, de los cuales 79 registros forman parte del clúster c_som_1_1, 98 registros conforman el clúster c_som_1_2, 128 forman el clúster c_som_2_1 y 61 del clúster c_som_2_2.

El algoritmo Kohonen SOM agrupó los clusters basándose en la similitud de los valores de sus atributos. Los datos de un mismo grupo presentan características comunes, pero a su vez, los objetos entre los grupos deben ser diferentes.

- Descubrimiento del comportamiento del delito

Para descubrir el comportamiento del delito, se establecen los atributos para la ejecución del algoritmo C4.5. El atributo clase se define como la variable grupo generada por el algoritmo Kohonen SOM en la formación de clusters, y se seleccionaron como los atributos de entrada: dia_m, mes_r, horas_r, delito_descrip, tipo_lugar, clase_arma, elemento_sustraido y tipo_ataque

La aplicación del algoritmo C4.5 proporcionó un total de 17 reglas que caracterizan a los grupos identificados. A partir de las reglas generadas, se elaboró una primera interpretación de los clusters resultantes:

- Clúster c_som_1_1: Caracterizado por delitos mayoritariamente cometidos por medio del forcejeo o arrebato. Incluye casos en la vía pública o en un domicilio particular. Registra mayor actividad delictiva los días miércoles, jueves y viernes, durante los meses de marzo, abril, mayo y junio.
- Clúster c_som_1_2: En principio se trataría de delitos a través del forcejeo o arrebato y con arma blanca. Incluye casos de ocurrencia en la vía pública o en un domicilio particular. Además, se observa la característica que los delitos ocurrieron durante los días sábado, domingo, lunes o martes, en los meses de marzo, abril, mayo y junio.
- Clúster c_som_2_1: Es el grupo o clúster que más delitos agrupa. Se determinó que el delito aconteció en algunos casos en forma de ataque brutal y en otros se determinó que no existió ningún tipo de ataque. Predomina el uso de arma fuego y en otros casos, se determinó que no existió ningún elemento de ataque, e incluye lugares del hecho como en la vía pública, en un comercio, en el interior de un rodado, o en un domicilio particular. Este grupo está caracterizado por el robo objetos personales, motocicletas, de tipo domiciliario, vehículos y dinero. No obstante, en otros casos también se observa que no hubo elemento sustraído.
- Clúster c_som_2_2: Este grupo a diferencia del resto, posee mayor descripción de las características de los delitos. Particularmente, los delitos ocurrieron con ataque brutal a las víctimas y en otros casos, no existió ningún tipo de ataque. Se detectó como el arma de fuego y arma blanca, los elementos predominantes en este grupo, así mismo, también resultaron casos en donde no se registraron elementos de ataque. Este grupo incluye casos de ocurrencia en vía pública o en algún domicilio particular, en horarios de mañana o siesta. Asimismo, los meses que caracterizan a este grupo son enero y febrero. Los elementos sustraídos durante el delito que definen a este grupo son de tipo objetos personales, motocicletas, de tipo domiciliario, vehículos y dinero. En otros casos de igual forma se observa que no hubo elemento sustraído.

Ponderación de atributos con mayor incidencia en el comportamiento de delitos

Se definen a continuación, los atributos para la ejecución del algoritmo Naive Bayes. El atributo clase se identifica como la variable de predicción generada por el algoritmo C4.5. en el descubrimiento del comportamiento del delito, y se establecen los atributos de entrada: dia_m, mes_r, horas_r, delito_descrip, tipo_lugar, clase_arma, elemento_sustraido y tipo_ataque.

A través del estudio de las características del delito con mayor incidencia en la ocurrencia del hecho, se

presenta el análisis para cada clúster:

- Clúster c_som_1_1: La mayor cantidad de delitos ocurrieron por robo (94%). En la mayoría de los casos se trata de arrebatos (79,76%) y en la vía pública (90,48%). Existe un alto porcentaje de incidencia (84,52%) de la sustracción de algún objeto personal de la víctima. Se detectó al día viernes, como el día de la semana más frecuente (29,76%), y en cuanto a los valores de incidencia por mes, marzo y abril registraron mayor volumen de delitos (60,71%).
- Clúster c_som_1_2: Este clúster se caracteriza por la ocurrencia de robos en su totalidad por arrebato (79,76%) y de objetos personales. La mayoría de los casos ocurridos sucedieron en vía pública (90,48%) y con arma blanca (92,63%). Los hechos delictivos tienen mayor ocurrencia los días sábados y domingos, en el horario de siesta de 12:00 pm a 16:00 pm y de noche de 20:00 pm a 24:00 pm. Los robos tienen mayor ocurrencia en los meses de marzo y abril según los valores de incidencia.
- Clúster c_som_2_1: Este clúster está determinado por la distribución en cantidades iguales en la ocurrencia de ambos tipos de delitos (robo y hurto). El objeto sustraído con mayor incidencia con respecto al resto fue de tipo objeto personal (35%). De los lugares, el más frecuente fue en la vía pública (33,33%).
- Clúster c_som_2_2: Se analizó que los delitos de robo ocurrieron con mayor frecuencia en los meses de enero y febrero, los días sábados y en el horario de siesta de 12:00 pm a 16:00 pm. Este clúster se caracterizó por presentar mayores casos en la vía pública (95%), utilizando arma blanca (65,67%). Los robos ocurrieron en su totalidad por ataque brutal (52,24%). En cuanto a los elementos sustraídos los más frecuentes fueron motocicletas y objetos personales.

Finalizadas la ejecución de las técnicas de MD descriptas para: la formación de clusters, el descubrimiento del comportamiento de los delitos y en la ponderación de atributos con mayor incidencia en dicho comportamiento se obtuvo valiosa información. Ésta resume que: i] cuatro de los atributos más significativos de los clusters son: tipo de objeto sustraído (objeto personal), lugares y tipos de armas utilizadas durante el delito (vía pública y con arma blanca) y tipo de ataque más sufrido (arrebato).

Los cluster c_som_1_2 y c_som_2_2, permitieron identificar días de mayor porcentaje de ocurrencia de delitos (viernes, sábados y domingos) y horarios más frecuentes (horarios de siesta y de madrugada).

Para la representación espacial de los clusters generados se utilizó como complemento la generación de un mapa de calor, una herramienta capaz de detectar las zonas con mayor concentración de puntos sobre una determinada ubicación geográfica.

Identificadas las zonas, se analizó la modalidad del clúster c_som_1_1 presente en ellas, con el objetivo de caracterizar los delitos de acuerdo a los resultados derivados de la aplicación de los algoritmos de minería de datos. Los barrios que agrupan estos puntos: Barrio 1, Barrio 2, Barrio 3, Barrio 4, Barrio 5, Barrio 6 y Barrio 7. La modalidad de delitos para este clúster se caracterizó por el uso de armas y mayoritariamente de tipo blancas, ocurrieron en la vía pública y a través del arrebato. Los días viernes se registraron mayor cantidad de hechos delictivos en el horario de siesta de 12:00 pm a 16:00 pm., detectándose los robos de objetos personales.

4. Conclusiones

El trabajo describe una integración entre tecnología GIS y minería de datos: en particular se aplicaron los algoritmos Kohonem SOM, C4.5 Y Naive Bayes. Los resultados obtenidos muestran la buena calidad de la información generada que innovaría en el modo de apoyar procesos decisorios.

Los mapas delictivos creados corresponden a la relación entre el delito ocurrido y un espacio geográfico y tiempo determinado. Estas representaciones gráficas aportan a la identificación y localización del delito con mayor frecuencia de ocurrencia.

En este objetivo de minería de datos, el análisis caracteriza el delito de acuerdo con variables como: la modalidad, el empleo de armas, los objetos sustraídos, lugar de ocurrencia, el rango horario, los días de la semana y los meses con más delitos ocurridos.

De acuerdo con la modalidad del delito estudiada en cada uno de los clusters, el arrebato es la constante en los cuatro grupos, el mayor número se registra en la zona de la vía pública. Los objetos mayormente sustraídos son de tipo personal, y predomina el uso de las armas blancas utilizadas para cometer los delitos.

Los delincuentes prefieren actuar los días viernes, sábados y domingos. En cuanto al rango horario, la tendencia de los hechos delictivos se registró entre las 12 p.m. y 16 p.m.

Respecto a los meses en los cuales se presentaron más delitos, los clusters coinciden en detectar a los meses de marzo y abril. Si bien este resultado no ofrece mayor información, como líneas futuras se podrían aplicar estas técnicas de minería de datos para los años siguientes y los resultados, y así conocer la influencia de delito por época o por estaciones del año.

Cómo citar este artículo / How to cite this paper

Flores, L. E.; Mariño, S. I.; Martins, S. (2019). Modelado y simulación de robos y hurtos basados en redes SOM, TDIDT y Bayesianas. Un caso de estudio. *International Journal of Information Systems and Software Engineering for Big Companies (IJISEBC),* 6(2), 81-87. (www.ijisebc.com)

Referencias

Anoopkumar, M.; Rahman, A. M. Z. (2016). A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE) (pp. 122-133). IEEE.

Baghdadi, N.; Mallet, C.; Zribi, M. (2018). QGIS and Generic Tools. John Wiley & Sons.

Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, 16.

García-Martínez, R.; Britos, P.; Rodríguez, D. (2013). Information mining processes based on intelligent systems. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 402-410). Springer, Berlin, Heidelberg.

McCue, C. (2014). Data mining and predictive analysis: Intelligence gathering and crime analysis. Butterworth-Heinemann.

Rakotomalala, R. (2004). TANAGRA project. (http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra)