IDENTIFICACIÓN Y DETECCION DE PATRONES DELICTIVOS BASADA EN MINERIA DE DATOS

Perversi, I.[1], Valenga, F.[2], Fernández, E.[3,4], Britos P.[3,4], García-Martínez, R.[3,4]

[1] Departamento de Ingeniería Industrial. ITBA
[2] Facultad de Informática Ciencias de la Comunicación y Técnicas Especiales. UM
[3] Centro de Ingeniería de Software e Ingeniería del Conocimiento. Escuela de Postgrado. ITBA
[4] Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. UBA

{enfernan, pbritos, rgm}@itba.edu.ar

RESUMEN

En esta comunicación se describen resultados preliminares de la línea de investigación sobre el uso de minería de datos aplicadas a la identificación y detección de patrones delictivos, analizando los homicidios dolosos cometidos en la República Argentina.

1. INTRODUCCIÓN

A partir de la crisis de finales de 2001, Argentina se vio afectada por una creciente ola de inseguridad caracterizada por un aumento en los índices delictivos y los niveles de violencia. Esta situación fue más profunda en los principales centros urbanos y llevó a tomar acciones coordinadas a nivel nacional tendientes a prevenir el delito. Una de estas medidas fue la creación del Sistema de Alerta Temprana (SAT) por parte del Ministerio de Justicia y Derechos Humanos. En el plano internacional, los ataques terroristas del 11 de septiembre han aumentado significativamente la preocupación por la seguridad interna en EEUU. Las agencias de inteligencia como la CIA o el FBI procesan y analizan información activamente en busca de actividad terrorista [Chen *et al*, 2004].

El análisis de los registros criminales es fundamental en la prevención del delito. Entre otras cosas, porque permite el diseño de políticas y planes de prevención efectivos. En Argentina este tipo de análisis se ha realizado históricamente mediante herramientas estadísticas descriptivas o deductivas, considerando fundamentalmente variables y relaciones primarias. Sin embargo, muchas veces la estadística descriptiva clásica no refleja la verdadera interrelación de las variables y por lo tanto, el problema real. Este contexto requiere un tratamiento estadístico más complejo que nos obliga a evolucionar en el análisis de información criminal.

En general, 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 [Kantardzic, 2002]. Por esta razón, en muchos casos, los registros almacenados son demasiado grandes o complejos como para analizar [Kantardzic, 2002] y superan el alcance de la estadística [Hand, 1997]. La Minería de Datos (*Data Mining*) es un proceso iterativo de búsqueda de información no trivial en grandes volúmenes de datos [Kantardzic, 2002]. Busca generar información similar a la que podría generar un experto humano: patrones, asociaciones, cambios, anomalías y estructuras significativas [Ochoa, 2004].

En el caso de la inteligencia criminal, la gran cantidad de información y de variables intervienientes justifican el uso de herramientas más potentes que la estadística convencional que permitan determinar relaciones multivariantes subyacentes. La minería de datos aplicada a la inteligencia criminal es un campo bastante nuevo y ha tenido un gran impulso en los últimos años en EEUU [Chen *et al*, 2003].

2. ESTADO DE LA CUESTIÓN

2.1. Agrupación de Datos

La agrupación o el clustering 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. El clustering identifica clusters, o regiones densamente pobladas, de acuerdo a alguna medida de distancia, en un gran conjunto de datos multidimensional [Chen et al., 1996]. El clustering se basa en maximizar la similitud de las instancias en cada cluster y minimizar la similitud entre clusters [Han & Kamber, 2001]. K-Means [Britos *et al*.,,2005] en un método particional de clustering 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.

2.2. Clasificación de Datos

ID3 es un sistema típico de construcción de árboles de decisión, el cual adopta una estrategia de arriba hacia abajo e inspecciona solo una parte del espacio de búsqueda. ID-3 garantiza que será encontrado un árbol simple, pero no necesariamente el más simple. ID-3 utiliza la teoría de la información para minimizar la cantidad de pruebas para clasificar un objeto. Una heurística selecciona el atributo que provee la mayor ganancia de la información. Una extensión a ID3, C4.5 [Quinlan, 1993] extiende el dominio de clasificación de atributos categóricos a numéricos. J48 es una implementación mejorada del algoritmo de árboles de decisión C4.5. El algoritmo J48 funciona bien con atributos nominales y numéricos. Un paso importante en la construcción del árbol de decisión es la poda, la cual elimina las ramas no necesarias, resultando en una clasificación más rápida y una mejora en la precisión de la clasificación de datos [Han & Kamber, 2001].

3. DEFINICIÓN DEL PROBLEMA

En la actualidad el SAT (Sistema de Alerta Temprana) se encuentra implementado a en todas las provincias reportando información de los hechos delictivos ocurridos en todo el país. Esta información esta siendo tratada a través de análisis estadístico sin hacer uso de técnicas ni herramientas de minería de datos. Lo cual implica que no se está haciendo un verdadero aprovechamiento de los datos obtenidos.
En este contexto resulta de interés explorar el uso de minería de datos basada en sistemas inteligentes en el proceso de identificación y detección de patrones delictivos comenzando con el análisis de homicidios dolosos cometidos en la República Argentina.

4. ABORDAJE DEL PROBLEMA

El problema de identificación y detección de patrones de homicidios dolosos cometidos en la República Argentina se abordo con la siguiente estrategia:

1. Clusterizar los datos relevantes homicidios dolosos cometidos.
2. Analizar los cluster obtenidos y validarlos con los usuarios
3. Aplicar algoritmos de inducción a cada cluster para encontrar explicaciones descriptivas del comportamiento que subyace a la pertenencia a los mismos

4.1 Estado de Avance

A la fecha se ha logrado identificar los atributos más significativos del cubo de datos y llegar a convalidar los resultados provenientes del proceso de agrupación (clustering) con los usuarios. Se encuentra en vías de desarrollo la aplicación de técnicas de inducción para explicar el comportamiento de los distintos grupos (clusters).

4.2. Descripción del Cubo de datos

Se analizaron 1810 registros de la base de datos "Homicidios Dolosos" correspondientes a la totalidad de hechos registrados durante 2005, provenientes del SAT. Cuyos atributos se describen a continuación en la tabla 1:

Provincia	Departamento	Día del mes	Mes	Día de la semana
Hora	Lugar	Arma	Otro delito	

Tabla 1. Atributos del Cubo de datos

4.3. Resultados del proceso de Agrupamiento

4.3.1. Centroides

A continuación, en la tabla 2, se describen los centroides obtenidos:

		Atributos categóricos (modas)			**Atributos continuos (medias)**			
	Cant. (%)	**Lugar**	**Arma**	**Otro Delito**	**Hora**	**Día Semana**	**Día Mes**	**Mes**
Cluster 0	21%	Domicilio Particular	de Fuego	Robo	15	Jueves	16	6
Cluster 1	53%	Vía Pública	de Fuego	No Hubo	12	Martes	15	6
Cluster 2	26%	Domicilio Particular	Ninguna	No Hubo	8	Miercoles	15	6
General	100%	Vía Pública	de Fuego	No Hubo	11	Miercoles	15	6

Tabla 2. Centroides

4.3.2. Interpretación de los Cluster

Cluster 0 (21%): esta caracterizado por homicidios mayoritariamente en ocasión de robo y con arma de fuego. La hora difiere significativamente de la media global, con una tendencia hacia la noche (antes de las 24hs). En principio diremos que se trata de *"homicidios en ocasión de robo"*.

Cluster 1 (53%): es el que más registros agrupa y el más parecido a la media global. Esta caracterizado por homicidios mayoritariamente en la vía publica con arma de fuego y sin la existencia de otro delito. Se podrían interpretar como *"homicidios en ocasión de riña o ajuste de cuentas"*.

Cluster 2 (26%): es el más particular de los clusters, ya que la mayoría de sus registros presentan casos de homicidios sin arma. La hora difiere de la media global con una tendencia hacia la madrugada (después de las 24hs). Los denominaremos *"homicidios en ocasión de emoción violenta"*.

4.3.3. Gráficos de Barras

La distribución de los clusters entre las variables de los distintos atributos permite comprender el nivel de significancia de los mismos (ver figura 1). En este caso, si los clusters fueran irrelevantes, esperaríamos encontrar una proporción aproximada de 50% rojo (cluster 1); 25% azul (cluster 0) y 25% turquesa (cluster 2) en cada variable de cada atributo. Si bien en algunos atributos esta proporción se cumple (*día mes*, *mes*) en otros existen interacciones significativas (por ejemplo cluster 2 con *ninguna arma* y *domicilio particular*).

4.3.4. Gráficos de dispersión

En las siguientes subsecciones se describen los cluster en base a dos de los atributos mas representativos

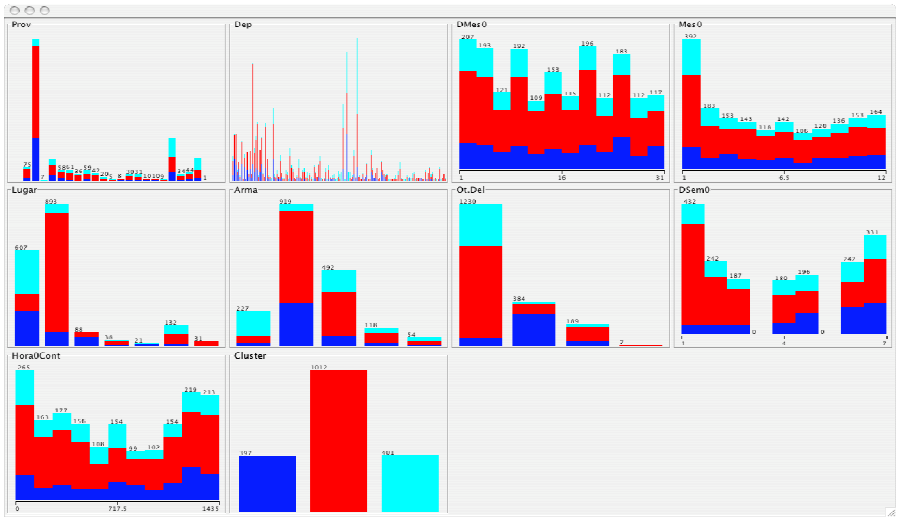

Fig. 1. Distribución de Clusters

Distribución de los clusters según el atributo lugar

Mientras el cluster 2 esta muy concentrado en domicilio particular y el cluster 1 en vía pública, el cluster 0 se encuentra distribuido más homogéneamente [Figura 5.2]. Si bien este último presenta la mayoría de registros en domicilio particular, tiene una alta proporción de homicidios en comercios respecto a los otros clusters (ver figura 2).

Distribución de los clusters según el atributo arma

El cluster 1 y el cluster 0 presentan una distribución similar (ver figura 3), con una alta concentración en arma de fuego, seguida por arma blanca y prácticamente muy pocos casos sin arma [Figura 5.3]. En contraposición el cluster 2 presenta muy pocos casos con arma de fuego (una proporción muy baja respecto a la proporción global) y muchos casos sin arma (una proporción muy alta respecto a la proporción global).

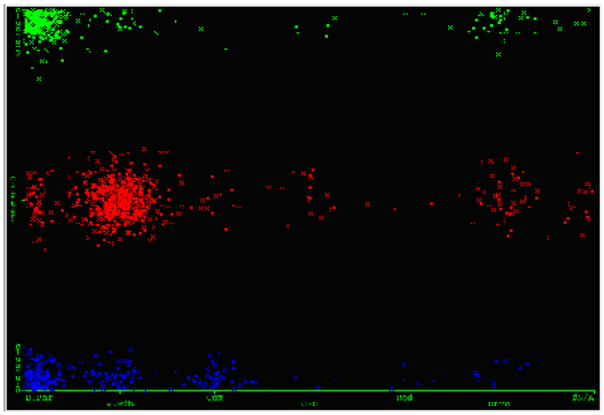

Fig. 2. Distribución según atributo lugar

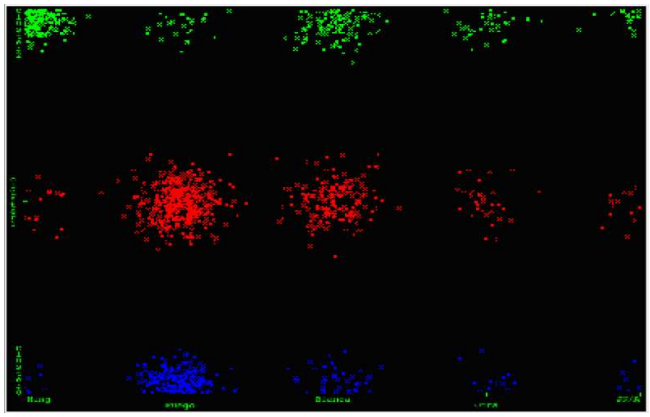

Fig. 3. Distribución según atributo Arma

Distribución de los clusters según el atributo lugar y arma

En la figura 4, se muestra el agrupamiento con base en la combinación conjunta de los atributos Lugar y Arma. Se observa: [a] existe una fuerte interacción entre *domicilio particular*, *ninguna arma* y cluster 2. En un nivel más general podríamos interpretar al cluster 2 como homicidios en

domicilio particular donde el arma *no es arma de fuego*, [b] se observa interacción entre *vía pública*, *arma de fuego*, y cluster 1. En un nivel más general podríamos interpretar al cluster 1 como homicidios en la *vía pública* con arma y [c] existe interacción entre *domicilio particular*, *arma de fuego* y cluster 0.

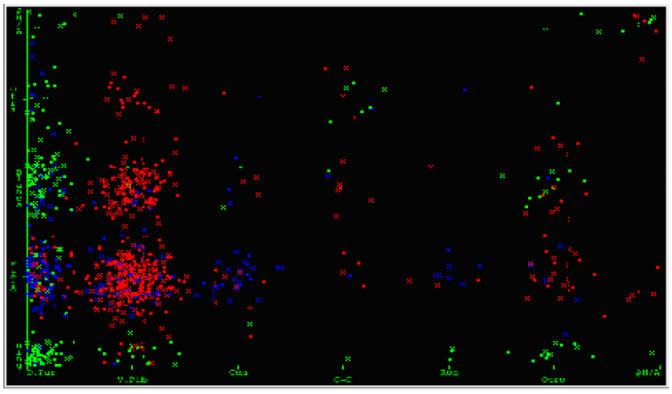

Fig. 4. Distribución del Cluster según los atributos Lugar y Arma

5. CONCLUSIONES

Existe información a partir de la cual es posible desarrollar un proyecto de Minería de Datos a gran escala para ayudar a la generación de políticas criminales en la República Argentina. Los conocimientos descubiertos como resultado de este proceso sirven para: [a] proporcionar una justificación sustentada en los datos disponibles de los conceptos preexistentes y [b] la detección de piezas de conocimiento sobre el dominio no identificable mediante otros métodos.
Se propone continuar con el proyecto: [a] **a**plicando técnicas de inducción para explicar en mayor detalle los cluster identificados y [b] ampliar el análisis a otros ámbitos y tipos de hechos (por ejemplo: homicidios dolosos causados por accidentes de tránsito).

6. FORMACIÓN DE RECURSOS HUMANOS

En la línea de investigación cuyos resultados parciales se reportan en esta comunicación, se encuentran trabajando dos tesistas de grado en ingeniería informática y tres investigadores formados.

7. AGRADECIMIENTOS

Los autores desean agradecer a la Secretaría de Política Criminal de la Nación por el apoyo que proporciona a este proyecto de investigación.

8. REFERENCIAS

Britos, P., Hossian, A., García-Martinez, R. y Sierra, E. 2005. *Minería de Datos Basada en Sistemas Inteligentes*. Nueva Librería.

Chen, H., W. Chung, J. Xu, G. Wang, Y. Qin, M. Chau, 2004. *Crime Data Mining: A General Framework and Some Examples*. IEEE Computer Society, vol. 37, no. 4, pp. 50-56.

Chen, M., Han, J., 1996. *Data mining: An overview from database perspective*. IEEE Transactions on Knowledge and Data Eng.,

Han, J., Kamber, M. 2001. *Data mining: Concepts and techniques*. Morgan Kauffmann Publishers,

Hand, D. J., 1997. *Data Mining: Statistics and More?*. The American Statistician.

Kantardzic, M. 2002. *Data Mining: Concepts, models, methods and algorithms.* Wiley-IEEE Press.

Ochoa, M. A. 2004. *Herramientas Inteligentes para la Explotación de Información*. Trabajo Final: Especialidad en Ingeniería en Sistemas Expertos, Instituto Tecnológico de Buenos Aires (ITBA).

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.