IDENTIFICACION DE PATRONES CARACTERISTICOS DE LA POBLACION CARCELARIA MEDIANTE MINERÍA DE DATOS

Gutiérrez Rüegg, P., Merlino, H., Rancan, C., Procopio, C., Rodríguez, D., Britos, P., García-Martínez, R.

Departamento de Ingeniería Industrial. Instituto Tecnológico de Buenos Aires
Centro de Ingeniería del Software e Ingeniería del Conocimiento. Instituto Tecnológico de Buenos Aires
Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. Universidad de Buenos Aires

{hmerlino, crancan, cprocopio, drodrigu, pbritos, rgm}@itba.edu.ar

RESUMEN

La situación penitenciaria muestra números preocupantes en sobrepoblación, hacinamiento, promiscuidad, avance del sida, violencia, entre otros. Estos incrementos y cuestiones alarmantes han instalado el tema como uno de los puntos inevitables de estudio dentro de la política criminal. En este contexto, en este trabajo se presentan resultados preliminares sobre el uso de minería de datos aplicada a la población carcelaria en Argentina con fines de prevención de delitos.

1. INTRODUCCION

La situación penitenciaria representa desde hace ya varios años un tema de especial preocupación en la Argentina. El crecimiento en las tasas de delitos registrado a lo largo de la década del noventa fue acompañado por un fuerte aumento en la población privada de libertad. Entre los años 1997 y 2006 dicha población aumentó alrededor del 83%, afectado profundamente por la crisis político social vivida entre los años 2001 y 2002, llegando al pico de cantidad de presos en el año 2005. En lo que refiere al régimen penitenciario en Argentina, las estadísticas muestran números realmente preocupantes: la sobrepoblación, el hacinamiento, la promiscuidad, el avance del sida, la violencia, son algunos de los temas que se hacen presentes cuando se habla de sociedades carcelarias en Argentina. Se cree, por lo tanto, que sin una política criminal eficiente y reparadora será imposible bajar en el mediano plazo el índice de delitos existente. Estos incrementos y cuestiones alarmantes han instalado el tema como uno de los puntos inevitables de estudio dentro de la política criminal.
La problemática planteada lleva a la necesidad de contar con herramientas que permitan desarrollar un diagnóstico válido sobre el estado actual de la cuestión carcelaria. El Sistema Nacional de Estadísticas sobre Ejecución de la Pena (SNEEP) es un aporte en tal sentido. Este sistema de información fue implementado en el año 2002 por la Dirección Nacional de Política Criminal con el fin de contar con información periódica y uniforme acerca de la población penal privada de libertad en la República Argentina. Sin embargo, se cree que con el solo contar con estadísticas vinculadas a los registros no es suficiente para poder tomar decisiones completas y correctas en lo que a política criminal respecta. Es necesario un tratamiento de la información estadística más complejo, aún más cuando se trabajan con base de datos de más de 50000 registros, como es el caso de la población carcelaria en Argentina. El objetivo primario de estas bases de datos es, como su nombre indica, almacenar grandes cantidades de datos organizados siguiendo un determinado esquema o modelo de datos que facilite su almacenamiento, recuperación y modificación, pero no así su posterior uso o análisis. En muchos casos los registros almacenados son demasiados grandes y complejos como para analizar [Kantardzic, 2003]. Una posible herramienta a utilizar para tratar los grandes volúmenes de información almacenados en bases de datos es la técnica de *Minería de Datos*. La minería de datos representa la posibilidad de buscar exhaustivamente dentro de un gran volumen de datos información y conocimiento que pueden resultar de mucho valor similar a la que podría generar un experto humano: patrones, cambios, anomalías y estructuras significativas [Britos *et al*, 2005].

2. ESTADO DE LA CUESTIÓN

2.1. Clustering o Agrupamiento de los Datos

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. Este tipo de algoritmo se realiza en forma no supervisada ya que no se saben de antemano las clases del conjunto de datos de entrenamiento. El clustering identifica regiones densamente pobladas, de acuerdo a alguna medida de distancia, en un gran conjunto de datos multidimensional [Chen & Han, 1996]. El análisis de clusters se basa en maximizar la similitud de las instancias en cada cluster y minimizar la similitud entre clusters [Han & Lamber, 2001]. Es utilizado en numerosas aplicaciones tales como reconocimiento de patrones, análisis de datos, procesamiento de imágenes e investigaciones de mercado. Como función de la *minería de datos*, el análisis de clusters puede ser utilizado como una herramienta independiente para obtener una visión de la distribución de los datos, para observar las características de cada cluster y enfocar un análisis más exhaustivo hacia un grupo o cluster determinado. Alternativamente, puede servir como un paso del preprocesamiento de los datos para otros algoritmos, como por ejemplo, el de clasificaciones en el cual se trabajaría luego sobre los clusters originados.

2.2. Clasificación de los Datos. Algoritmos de Inducción

Los algoritmos de clasificación se utilizan para clasificar un conjunto de datos basado en los valores de sus atributos [Servente & García Martínez, 2002]. El objetivo de la clasificación es analizar los datos de entrenamiento y, mediante un método supervisado, desarrollar una descripción o un modelo para cada clase utilizando las características disponibles en los datos Los algoritmos más utilizados para la clasificación son los algoritmos de inducción. Aún cuando existen varios enfoques para los algoritmos de inducción, se trabajará con aquellos que generan árboles de decisión conocida como la familia TDIT (*Top Down Induction Trees*). En particular se utilizará el *Chi Squared Automatic Interaction Detection*: CHAID [Hartigan, 1975]. El mismo se sirve de la Prueba de la x^2 (chi squared test) para determinar si se debe continuar con la ramificación y, en caso afirmativo, qué variables independientes usar. El modelo utiliza el algoritmo CHAID para dividir en grupos los registros que presenten la misma probabilidad de resultado, basándose en los valores de las variables independientes. El algoritmo parte de un nodo raíz y se va bifurcando en nodos descendientes hasta llegar a los nodos hoja, donde finaliza la ramificación.
Entre otros importantes algoritmos de árboles de decisión, se puede destaca el ID3 [Quinlan, 1986] y su extensión C4.5 [Quinlan, 1993]. El J48 es una implementación mejorada del algoritmo C4.5, funcionando bien tanto con atributos nominales como numéricos. Cabe destacar, que a modo de prueba para el estudio en cuestión, se utilizó tanto el CHAID como el J48 obteniéndose resultados muy similares con ambos algoritmos.

3. DESCRIPCIÓN DEL PROBLEMA

Los incrementos en la población carcelaria hacen suponer que las tasas de delitos deberían haber bajado, sin embargo ocurre lo contrario. Actualmente el SNEEP solo publica estadísticas sin tratamiento alguno desaprovechando toda la información que proporcionan los datos. Se buscar detectar relaciones entre la variables realizando un estudio que diagnostique y ayude a proponer alternativas para prevenir que una persona cometa un delito.
En este contexto, el objetivo del trabajo es el de caracterizar a la población carcelaria mediante técnicas de minería de datos, esperando encontrar relaciones subyacentes en los datos que no pueden identificarse mediante un tratamiento estadístico clásico.

4. ABORDAJE DEL PROBLEMA

A los fines de caracterizar a la población carcelaria y encontrar conductas y patrones de comportamiento en los reclusos, se prosiguió a encarar la problemática de la siguiente manera:

1. Proceso de Clustering utilizando atributos significativos de los presos.
2. Análisis y validación de los clusters obtenidos con especialistas
3. Aplicación de algoritmos de inducción a cada cluster para la identificación de reglas de decisión que ayuden a explicar la composición de cada grupo.

4.1. Estado de Avance

Se ha avanzado con la identificación de los atributos más importantes de la población carcelaria para su posterior proceso de clustering. El estudio se encuentra en la etapa de validación de los clusters con especialistas en el tema (fiscales penales, jueces y sociólogos). A su vez, al momento, se han realizado a modo de prueba corridas de clasificación de los datos. Este último paso aún se encuentra en su etapa de desarrollo y validación.

4.2. Descripción del Dataset

Se analizaron 50408 registros de presos masculinos pertenecientes a la base de datos "Censo Población Carcelaria" provenientes del SNEEP. Una vez que se realizada la fase de recolección, exploración y limpieza de los datos iniciales, se preparó el siguiente conjunto de datos (40928 registros) con sus respectivos atributos.

Edad	Estado Civil	Nivel de Instrucción	Situación Laboral
Lugar de Residencia	Capacitación Laboral	Delito Cometido	Reincidencia

Tabla 1. Atributos del Dataset

4.3. Resultados del Clustering de los Datos

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Delito Cometido	Contra la Propiedad	Contra las Personas	Contra la Propiedad	Violación / Drogas
Nivel de Instrucción	Primario Completo	Primario Completo	Primario Completo	Primario Completo
Ultima Situacion Laboral	Tiempo Parcial	Desocupado	Desocupado	Tiempo Parcial
Capacitación Laboral	Oficio	Ni Oficio ni Profesión	Ni Oficio ni Profesión	Oficio / Profesión
Estado Civil	Soltero	Soltero	Soltero	Casado
Lugar de Residencia	Urbana	Urbana	Urbana	Urbana
Edad Promedio	31	34	27	43
Total	16849 (41%)	6513 (16 %)	14662 (36%)	2904 (7%)

Tabla 2. Centroides obtenidos mediante el clustering.

En primer lugar, y antes de analizar cada cluster por separado, se observa un *nivel de instrucción* muy pobre, en donde primario completo e incompleto agrupan al 75% de los casos. Los mismo ocurre con el atributo *último lugar de residencia*, donde aproximadamente el 90% de las instancias corresponden a Urbana.

4.3.1. Primera Interpretación de los Clusters

Cluster 0 (41%): el que más registros agrupa**,** podría tratarse de personas, que aun cuando cometieron **delito contra la propiedad (mayormente robo)** no hay patrones que indiquen que lo hayan hecho por una necesidad marcada. Son personas que trabajan parcialmente y tenían algún

oficio. Generalmente reincidentes, y que sus salarios les alcanza para lo básico, teniendo que salir a robar para complementar sus necesidades y la de sus hijos.

Cluster 1 (16%): cluster que agrupa a los presos que delinquieron **contra las personas**. Generalmente desocupados y sin oficio ni profesión, estaría caracterizado por un lado por las personas que cometieron homicidio en ocasión de robo. A su vez, podría decirse que al ser personas totalmente inactivas a una edad en donde conseguir trabajo se les hace casi imposible, pueden haber llegado a delinquir contra las personas en una reacción de emoción violenta.

Cluster 2 (36%): segundo grupo en importancia, agrupa a los jóvenes que no tienen estudios, trabajo, ni profesión alguna. En un principio, este estado de exclusión de un régimen laboral los llevaría a **robar y/o hurtar** como única salida para poder subsistir. Importante tener en cuenta que es uno de los clusters más preocupantes por tratarse de gente joven, en donde las drogas pesadas juegan un papel muy importante.

Cluster 3 (7%): cluster más difícil de interpretar, agrupa en mayor medida a las personas que cometieron delitos contra la **Integridad Sexual** y por quienes fueron procesados o condenados por **Estupefacientes**. Con un promedio de edad mayor a los 40 años y casados o en concubinato, no se observan patrones que los hayan llevado a delinquir por una necesidad especifica ya que se trata de personas con algún tipo de empleo de tiempo parcial o completo.

4.3.2. Gráficos de Barras

Se puede observar [figura 1] como se distribuyen significativamente las variables de los atributos en los distintos clusters. Si bien en los atributos nivel de instrucción y ultima residencia la distribución en los clusters es irrelevante, ya que se cumple la proporción 41% rojo (cluster 0) 16% gris (cluster 1) 36% turquesa (cluster 2) 7% azul (cluster 3), en los otros atributos se pueden encontrar interacciones significativas.

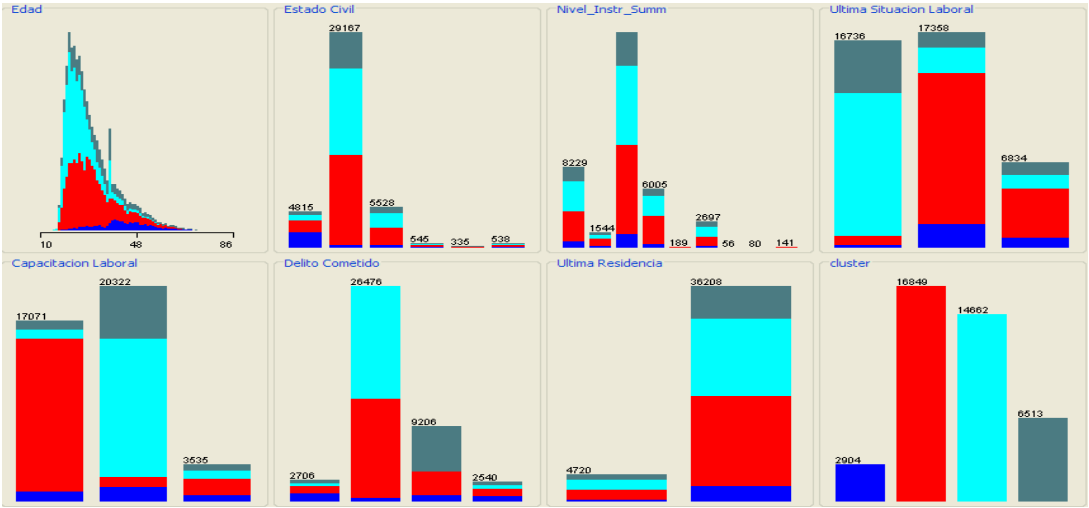

Figura 1. Distribución de los clusters entre las variables de los distintos atributos

4.3.3. Gráficos de dispersión

Tres de los atributos más significativos son el *delito cometido*, la *última situación laboral* y la *capacitación laboral***.** Tal como puede observarse en la figura 2, existe una interacción importante entre el cluster 2 (verde), no tiene oficio ni profesión y delito contra la propiedad. A su vez, en la figura 3 puede se aprecia que el cluster 2 esta caracterizado por personas desocupadas, mientras que el cluster 0 (rojo) concentra mayor cantidad de instancias en la situación de oficio de tiempo parcial/completo. El delito que caracteriza al cluster 0 es contra la propiedad [figura 2].

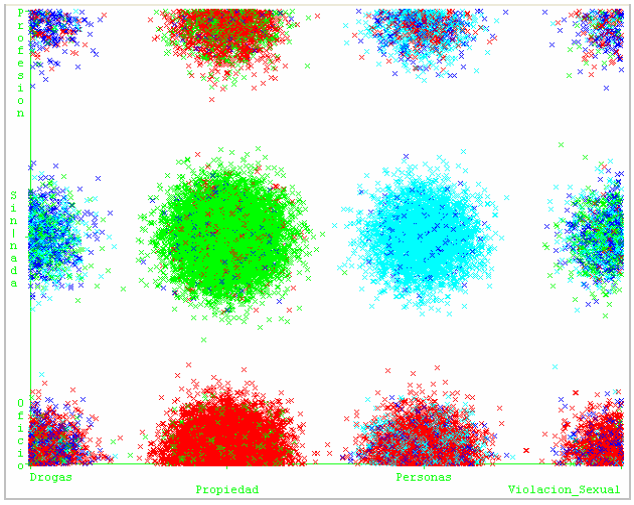

Figura 2. Distribución según Delito-Capacitación

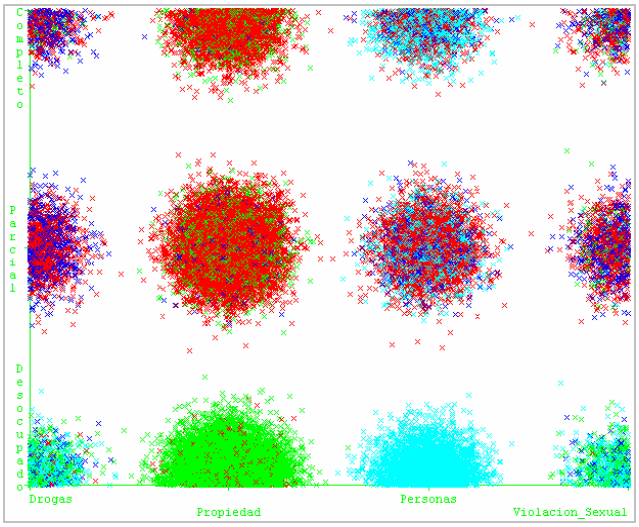

Figura 3. Distribución según Delito-Situación Laboral

En lo que respecta al cluster 1 (turquesa) se observa interacción con delito contra las personas y sin oficio ni profesión [figura 2], en su mayoría se trata de personas desocupadas [figura 3]. A su vez, el cluster 3 (azul) se distribuye en delitos contra la integridad sexual, estupefacientes y en menor medida delitos contra las personas [figura 2]. Generalmente se observan que son personas con trabajos de tiempo parcial o completo.

5. CONCLUSIONES

En primer lugar se destaca la factibilidad de aplicar modelos de minería de datos para el tratamiento de información relativa a poblaciones carcelarias. Se encontraron interacciones interesantes que no llegan a observarse a simple vista y que podrían ayudar a generar una política criminal reparadora intramuros y preventiva fuera de las cárceles.

Se continuará el proyecto de la siguiente manera: [a] aplicando técnicas de inducción para explicar más detalladamente los clusters formados [b] analizando en conjunto los resultados obtenidos durante el estudio relacionándolos con información recolectada a fin de obtener conclusiones que sirvan para la elaboración de programas en lo que a política criminal respecta.

6. FORMACIÓN DE RECURSOS HUMANOS

En la línea de investigación cuyos resultados parciales se reportan en esta comunicación, se encuentran trabajando un tesista de grado y tres investigadores en formación.

7. AGRADECIMIENTOS Y FINANCIAMIENTO

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-Martínez, R. y Sierra, E., 2005. *Minería de Datos Basada en Sistemas Inteligentes*. Editorial Nueva Librería. Buenos Aires. ISBN 987-1104-30-8.

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

Kantardzic, M., 2003. *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons. I

Hartigan, J.A., 1975. *Clustering algorithms*. John Wiley & Sons, New York.

Quinlan, J., 1993. *Programs for Machine Learning.* Morgan Kaufmann Publishers. Edición 1993.

Servente, M.; García-Martínez, R., 2002. *Algoritmos TDIDT Aplicados a la Minería Inteligente.* http://www.fi.uba.ar/laboratorios/lsi/R-ITBA-26-datamining.pdf. Vigente al 28-02-08.