Minería de datos para la determinación del grado de exclusión social

Data mining to determine the degree of social exclusion

[*]**Jorge Enrique Rodríguez Rodríguez**

Fecha de recepción: 23 de agosto de 2008
Fecha de aceptación: 5 de octubre de 2008

Resumen

Se muestra el proceso de aplicación de minería de datos en un problema específico relacionado con la determinación del grado de exclusión social, dado un conjunto de atributos. Para tal fin se emplea una red neuronal artificial con topología con conexión hacia delante. Este artículo es un avance parcial del proyecto de investigación "Desarrollo de herramientas para minería de datos - UDMiner".

Palabras clave: minería de datos, red neuronal, clasificación de datos, modelo predictivo.

* Magíster en Ingeniería de Sistemas. Especialista en Ingeniería de Software. Especialista en Diseño y Construcción de Soluciones Telemáticas. Ingeniero de Sistemas. Docente investigador de la Universidad Distrital Francisco José de Caldas. Director del grupo de investigación en Inteligencia Artificial de la misma Universidad. jrodri@udistrital.edu.co

Abstract

In this paper I show the development process of Data Mining in a specific issue related to the degree of social exclusion, given a set of features. For this purpose it uses a neural network with topology feedforward. This paper is presented as a partial advance of the research project "Development of Tools to Data Mining – UDMiner".

Key words: Data mining, neural network, data classification, predictive model.

Introducción

Al hablar de minería de datos es necesario hacer referencia a las áreas con las cuales tiene relación; la estadística tradicional y el análisis de datos son algunas de estas. Los métodos estadísticos y el análisis sobre los datos no proporcionan conocimiento como tal; debido a esto fue necesario fomentar una práctica más profunda para utilizar los datos y extraer beneficios de estos. La respuesta a estas necesidades y a muchas otras, como el almacenamiento de gran cantidad de datos y la necesidad de herramientas adecuadas e innovadoras que apoyen la toma de decisiones, está reflejada en una de las áreas de investigación más recientes, la minería de datos.

A continuación se dan algunas definiciones:

La minería de datos es la exploración de forma automática o semiautomática de grandes cantidades de datos para el descubrimiento de reglas y patrones [1].

La minería de datos es la búsqueda de nueva y valiosa información no trivial en grandes volúmenes de datos [2].

La minería de datos puede definirse como un proceso iterativo de detección y extracción de patrones a partir de grandes bases de datos: esto es modelo-reconocimiento [3].

La minería de datos es el análisis de un conjunto de datos para encontrar relaciones desconocidas y resumir los datos de nuevas formas entendibles para el minero [4].

En la práctica, los modelos para extraer patrones 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 la base de datos, llamadas variables independientes o predictivas. Los modelos descriptivos, en cambio, identifican patrones que explican o resumen los datos, es decir sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos [5].

En este documento se presenta la aplicación de minería de datos para el programa integral de formación laboral en áreas técnicas y en actividades de mejoramiento y mantenimiento del espacio público de la Unidad de Extensión, Facultad Tecnológica de la Uni-

versidad Distrital Francisco José de Caldas, siguiendo una metodología de minería de datos.

1. Aspectos básicos

1.1. Identificar el problema

A menudo en la Unidad de Extensión de la Facultad Tecnológica de la Universidad Distrital Francisco José de Caldas se desarrollan convenios o contratos de capacitación a personas de los estratos menos favorecidos. Uno de estos convenios es la Capacitación Laboral en Áreas Técnicas y en Actividades de Mejoramiento y Mantenimiento del Espacio Público, dirigido a jóvenes adscritos al Instituto Distrital para la Protección de la Niñez y la Juventud (Idipron). Este programa, además de la capacitación, pretende brindar a sus beneficiarios un cambio personal, económico, cultural y social. Puesto que dichos jóvenes se caracterizan por encontrarse en riesgo de exclusión social, por diferentes circunstancias, como la influencia del entorno de su familia y otros aspectos tenidos en cuenta por el psicólogo o trabajador social del programa o del Idipron, en el momento del ingreso del estudiante al programa se determina el riesgo de exclusión social que puede ser alto o bajo; el cual se utilizará a través del desarrollo del programa de capacitación para observar el proceso de cada uno de los jóvenes.

Para determinar el riesgo de exclusión social, además de lo mencionado anteriormente, las personas encargadas de este proceso se basan también en visitas domiciliarias (para determinar el lugar de residencia, la localidad, etc.), la edad de los jóvenes, su ocupación, entre otros factores.

De acuerdo con lo anterior, al momento del ingreso de un nuevo joven al programa o incluso al Idipron, se necesita hacer un estudio sobre este para determinar el riesgo de exclusión social; por tal razón, sería interesante y de mucha ayuda para las personas que realizan este estudio. Basado en lo anterior se aplicará la herramienta de minería de datos UDMiner con el fin de clasificar el riesgo de exclusión social para apoyar el proceso en mención.

1.2. ¿Es necesario el esfuerzo KDD[1]?

En el proceso de ingreso de jóvenes al programa o a Idipron, es necesario evaluar diferentes aspectos para determinar el riesgo de exclusión social de cada joven, proceso que implica un periodo considerable y un amplio conocimiento acerca de las variables involucradas en la determinación de exclusión social; por este motivo desarrollar un sistema de minería de datos que apoye este proceso es de gran beneficio para las personas encargadas de esta labor, pues basándose en los resultados del sistema se podría agilizar esta tarea y tomar mejores decisiones.

El esfuerzo KDD es necesario ya que, por medio de la extracción del conocimiento de la base de datos del programa de formación laboral, los expertos (psicólogo, orientador social, asistentes administrativos, coordinadores) encontrarán en la aplicación un apoyo para la toma de decisiones.

1 Descubrimiento de conocimiento a partir de grandes volúmenes de datos.

1.3. ¿Hay algún segmento que sea más interesante?

Debido a que se trata de realizar un sistema que apoye al psicólogo o trabajador social en la labor de calificar el riesgo de exclusión social de los jóvenes que ingresan al programa, además de otros aspectos que pueden ser complemento para la información almacenada de la Unidad de Extensión en cuanto al desarrollo del programa, se ha determinado que se debe aplicar minería de datos sobre la tabla (entidad) que posee la mayor cantidad de información y la más relevante que se pueda utilizar para tal fin, en este caso es la entidad Inscritos de la base de datos del Programa Integral de Formación para el Trabajo en Áreas Técnicas y en Actividades de Mejoramiento y Mantenimiento del Espacio Público, que posee la Unidad de Extensión de la Facultad Tecnológica.

1.4. Fuentes de datos

Existen dos posibles fuentes de datos: 1. La base de datos que posee la Unidad de Extensión de la Facultad Tecnológica en desarrollo del Programa Integral de Formación para el Trabajo en Áreas Técnicas y en Actividades de Mejoramiento y Mantenimiento del Espacio Público; y más específicamente la entidad Inscritos. 2. La base de datos que contiene la información general de los jóvenes pertenecientes al Idipron.

Se eligió como fuente de datos la base de datos que posee la Unidad de Extensión de la Facultad Tecnológica en desarrollo del Programa Integral de Formación para el Trabajo en Áreas Técnicas y en Actividades de Mejoramiento y Mantenimiento del Espacio Público; y más específicamente la entidad Inscritos.

Esta fuente de datos es válida, pues, aunque no contiene un número alto de registros como los podría contener la base de datos de Idipron, sí contiene información vital para la aplicación de minería de datos.

1.5. ¿Qué dicen los expertos?

Al plantear a los expertos (psicólogo, trabajador social, comunicadores sociales), directivos y otras personas pertenecientes al programa (administradores, profesores) y a las personas encargadas de calificar el riesgo de exclusión social de los jóvenes participantes del programa, la idea de implementar minería de datos que se utilizaría como apoyo en la realización de esta tarea, se observó buena aceptación ya que reduciría el tiempo para este proceso e incluso costos (personal, transportes, papelería), no solo para ellos sino también para el programa, pues, con la ayuda que proporcionaría la herramienta, no sería necesaria la participación de más de dos personas en este proceso y ya no sería necesario realizar la visita que se hace al lugar de residencia de los jóvenes; además, no se incurriría en gastos de papel utilizado para dejar el registro de dichas visitas. Lo anterior se suma a uno de los principales objetivos de la minería de datos: extraer patrones de conocimiento a partir de grandes volúmenes de datos.

1.6. ¿Qué es importante de acuerdo con la intuición y experiencia?

Para las directivas del programa de formación laboral, es importante que la información resultante luego del proceso de minería de datos sea clara y que brinde a las personas que la utilicen finalmente, como el psicólogo o el trabajador social, una ayuda verdadera que apoye las decisiones que se toman

y las calificaciones que se hacen no solo al momento del ingreso de jóvenes al proyecto sino también durante el desarrollo del programa y de futuros programas.

2. Preparación de datos

2.1. Identificar requerimientos de datos

Si se trata de obtener modelos en los cuales debe aparecer información general de los inscritos al programa, se debe establecer como objetivo minar la entidad donde aparezca la mayor cantidad de información y la más relevante, concerniente a este tema, y que se utilice para tomar decisiones.

La información que entrega la Unidad de Extensión se encuentra en la tabla Inscritos, la cual contiene el ID del registro, el nombre, los apellidos, el número de identificación, la edad, la dirección, la localidad, el nivel de escolaridad, la ocupación y el riesgo de exclusión social.

La Unidad de Extensión de la Facultad Tecnológica entregó una copia de la entidad Inscritos (tabla de la base de datos) exportada a formato de Excel 2000, debido a que no era posible acceder a una copia de la base de datos completa, ya que en esta, además de la información de los jóvenes también se encuentra información financiera del programa, la cual es de uso restringido.

Luego de esto se transformaron los datos al formato *.csv (delimitado por puntos y coma) donde se guarda únicamente el texto y los valores que aparecen en las celdas de la hoja de cálculo. Este formato garantiza que todas las filas y todos los caracteres de cada celda se almacenarán. Las columnas son separadas por un punto y coma (";") y cada fila se identifica por terminar en un retorno.

Exploración y limpieza de datos: se hizo una exploración sobre los datos en la cual se observaron algunas irregularidades y aspectos que se deben corregir para que el proceso de minería sea más efectivo.

• No existe homogeneidad en datos iguales del mismo campo; por ejemplo, en el campo localidad, el ítem Ciudad Bolívar podría aparecer como: 1. Ciudad Bolívar, 2. ciudad bolívar, 3. Ciudad_ Bolívar. Por lo cual se unificaron en los casos que son iguales pero están escritos de diferente forma.

• Los campos Estudia y Trabaja (ocupación) eran casillas de verificación, las cuales se remplazaron con las palabras sí y no, con el fin de que existiese claridad en la clasificación que obtenga el proceso de minería de datos.

A los datos se les realizó preprocesamiento utilizando la herramienta Preprocesar[2]; estos datos sufren una modificación dado que la herramienta en mención requiere de un formato preestablecido (tabla 1). Una vez almacenado el archivo que contiene estos datos (423 patrones), se procede a realizar el preprocesamiento, el cual se compone de: relleno de datos faltantes[3], selección de atributos[4], discretización, numerización y normalización[5] [6].

2 Esta herramienta está incluida en UDMiner, y se presenta como una primera fase dentro del proceso de minería de datos.

3 El relleno de valores faltantes se realiza a través del algoritmo EM (maximización de la esperanza).

4 La selección de atributos se lleva a cabo por medio de un árbol de inducción.

5 Se implementa la normalización de máximos y mínimos, más conocida como Normalización MAX-MIN.

Tabla 1. Una porción de los datos extraídos de la entidad Inscritos

Edad	Localidad	Estudia	Nivel escolar	Trabaja	Riesgo de exclusión social
14	Ciudad Bolívar	Sí	Décimo	No	Alto
14	Ciudad Bolívar	Sí	Noveno	Sí	Bajo
14	Ciudad Bolívar	Sí	Octavo	Sí	Bajo
14	Ciudad Bolívar	Sí	Octavo	No	Alto

3. Construcción del modelo

3.1. Crear el modelo

Los datos relevantes para el proceso de minería se encuentran en la entidad Inscritos de la base de datos que posee la Unidad de Extensión de la Facultad Tecnológica, creada en desarrollo del programa de capacitación, los cuales fueron previamente preprocesados para su posterior minado.

3.2. Escoger la mejor técnica

La tarea de minería de datos utilizada es la clasificación, ya que por medio de esta se puede predecir el riesgo de exclusión social basado en las características antes mencionadas; del mismo modo, se pueden establecer relaciones interesantes entre conjuntos de datos en una determinada clase de datos (riesgo de exclusión social). Para tal proceso de minería se utilizó una red neuronal artificial tipo *feedforward* junto con el algoritmo *backpropagation*[6].

3.3. Verificar el desempeño del modelo

Para verificar el desempeño del modelo, la red se entrena con 423 patrones (previamente preprocesados) y la prueba se lleva a cabo con 88 patrones diferentes a los de entrenamiento.

Porción de datos para preprocesar

```
%Extension
@nombre extension
@ATRIBUTO edad real
@ATRIBUTO localidad{antonio-nariño, bosa,chapinero,
ciudad-bolivar, engativa, fontibon, kennedy, martires,
puente-aranda, rafael-uribe-uribe, san-cristobal, soacha,
tunjuelito, usme}
@ATRIBUTO estudia {si, no}
@ATRIBUTO escolaridad {primero, tercero, cuarto,
quinto, sexto, septimo, octavo, noveno, decimo,
undecimo}
@ATRIBUTO trabaja {si, no}
@ATRIBUTO riesgo {bajo, alto}
```

6 Este algoritmo se basa en la regla delta generalizada; se suele implementar sobre redes neuronales artificiales *feedforward*.

```
@data
17,ciudad-bolivar,no,cuarto,si,alto
20,ciudad-bolivar,no,cuarto,no,alto
27,soacha,si,cuarto,si,bajo
16,bosa,si,decimo,no,alto
16,bosa,si,decimo,si,bajo
18,bosa,si,decimo,si,bajo
18,bosa,si,decimo,si,bajo
19,bosa,si,decimo,si,bajo
20,bosa,si,decimo,si,bajo
14,ciudad-bolivar,si,decimo,no,alto
15,ciudad-bolivar,si,decimo,si,bajo
17,ciudad-bolivar,si,decimo,si,bajo
17,ciudad-bolivar,si,decimo,si,bajo
17,ciudad-bolivar,si,decimo,si,bajo
```

En la tabla 2 se muestran las configuraciones de entrenamiento para las diferentes pruebas.

Porción de datos preprocesados

```
@nombre extension
@atributo estudia=si real
@atributo estudia=no real
@atributo estudia=nn real
@atributo trabaja=si real
@atributo trabaja=no real
@atributo trabaja=nn real
@atributo clase=bajo real
@atributo clase=alto real
```

```
@data
1,0,0,0,1,0,0,1
1,0,0,1,0,0,1,0
1,0,0,1,0,0,1,0
1,0,0,1,0,0,1,0
1,0,0,0,1,0,0,1
0,1,0,1,0,0,0,1
1,0,0,1,0,0,1,0
1,0,0,0,1,0,0,1
1,0,0,1,0,0,1,0
```

La cantidad de épocas se obtuvo a través de la experimentación, siendo este un valor apropiado para el proceso de minería, dado que se llegó a buenos resultados con un costo computacional bajo.

A continuación se muestran las matrices de confusión del entrenamiento (tabla 3) y la clasificación/predicción (tabla 4). Las demás se obvian pues, como se muestra en la tabla anterior, los resultados son los mismos (efectividad de clasificación del 100%).

Tabla 2. Configuración y resultados del entrenamiento (pruebas 1, 2, 3 y 4)

Configuración	Prueba 1	Prueba 2	Prueba 3	Prueba 4
Épocas	100	100	100	100
Neuronas ocultas	7	3	4	4
Capas ocultas	1	1	2	3
Neuronas de entrada	6 (posterior del preprocesamiento)	6 (posterior del preprocesamiento)	6 (posterior del preprocesamiento)	6 (posterior del preprocesamiento)
Neuronas de salida	2 clases	2 clases	2 clases	2 clases
Tasa de aprendizaje	0.3	0.3	0.3	0.3
Moméntum	0.2	0.2	0.2	0.2
Patrones de entrenamiento	335	335	335	335
Patrones de prueba	88	88	88	88
Efectividad del entrenamiento	100%	100%	100%	100%
Efectividad de la clasificación	100%	100%	100%	100%
Tiempo de entrenamiento	4 seg.	3 seg.	11 seg.	13 seg.

Tabla 3. Matriz de confusión - fase de entrenamiento

	C1	C2	Total	Éxito	Error
C1	200	0	200	100%	0%
C2	0	155	155	100%	0%
Total	200	155	355	100%	0%

Clases correctas: 355
Clases incorrectas: 0
Éxito del entrenamiento: 100%
Error del entrenamiento: 0%

Tabla 4. Matriz de confusión - fase de prueba (clasificación/predicción)

	C1	C2	Total	Éxito	Error
C1	47	0	47	100%	0%
C2	0	41	41	100%	0%
Total	47	41	88	100%	0%

Clases correctas: 88
Clases incorrectas: 0
Éxito de la prueba: 100%
Error de la prueba: 0%

4. Conclusiones

En este artículo se empleó una red neuronal artificial con conexión hacia adelante con el fin de determinar si una persona, dado un

conjunto de características, puede, o no, ser excluida socialmente. Luego de aplicar todo el proceso metodológico de minería de datos, se obtienen unos resultados que pueden utilizarse para mejorar las condiciones de vida de un grupo de personas. De igual forma se corrobora lo planteado por algunos autores cuando se afirma que las redes neuronales son una excelente técnica para clasificar datos.

Por otro lado, como conclusión específica del proyecto de investigación, se logra establecer un marco metodológico para minería de datos que puede ser utilizado para solucionar otros problemas. Del mismo modo, se establece que las redes neuronales son altamente efectivas en la solución del problema planteado.

5. Trabajos futuros

Se tiene previsto implementar otros algoritmos (redes bayesianas, algoritmos genéticos y métodos basados en casos) para clasificación de datos, los cuales serán probados y comparados en cuanto a efectividad y complejidad computacional con las redes neuronales.

Bibliografía

[1] Berry, M., Linoff, G. *Data mining techniques*. John Wiley & Sons. USA. 1997, p. 5.

[2] Kantardzic, Mehmed. *Data mining: Concepts, models, methods, and algorithms*. Wiley – Interscience. USA. 2001, p. 2.

[3] Mena, Jesús. *Data mining your website*. Digital Press. USA. 1999, p. 5.

[4] Hand, D., Mannila, H., Smyth, P. *Principles of data mining*. The MIT Press. USA. 2001, p. 1.

[5] Hernández, J., Ramírez, M., Ferri, C. *Introducción a la minería de datos*. Prentice Hall. España. 2004, p. 12.

[6] Barrera, H., Correa, J., Rodríguez, J. Prototipo de software para el preprocesamiento de datos "UD-Clear". IV Simposio Internacional de Sistemas de Información e Ingeniería de Software en la Sociedad del Conocimiento. Cartagena, Colombia, agosto de 2006, p. 165.