Minería de datos para la predicción de fraudes en tarjetas de crédito

Luis Felipe Wanumen Silvaz*

Fecha de recepción: octubre 15 de 2010
Fecha de aceptación: noviembre 5 de 2010

Resumen

En este artículo se expone el uso de la minería de datos a través de algoritmos de árboles de clasificación (J48) y reglas de asociación (a priori) para la posible detección de fraudes a nivel de tarjetas de crédito. Además, presenta una comparación de los resultados obtenidos con ambas técnicas y propone una serie de sugerencias para el desarrollo de este procedimiento usando minería de datos.

Palabras clave: minería de datos, árboles de clasificación, reglas de asociación, Algoritmo J48, Regla *a priori*.

* Ingeniero de sistemas, especialista en ingeniería de software de la Universidad Distrital Francisco José de Caldas. Docente de la facultad tecnológica. lwanumen@udistrital.edu.co

Abstract:
In this article describes the use of data mining algorithms through classification trees (specifically, J48) and rules of association (a priori) for the possible detection of fraud in credit cards. In addition, a comparison of the results obtained with both techniques, and proposes a number of suggestions for the development of this procedure using Data Mining.

Key words:
Data mining, skimming, credit card fraud, classification trees, association rules, algorithm J48, Rule *priori.*

Introducción

Después de los ataques del 11 de septiembre de 2001, agencias como la CIA y el FBI incrementaron sus bloques de inteligenca con un propósito principal: hallar información relacionada con los grupos terroristas. Por su parte, las técnicas más utilizadas en este proceso son:

1. Técnicas geográfico-visuales para detección de zonas calientes [1].
2. Standard Deviation Ellipses, mediante la cual pueden delimitarse agrupaciones de hechos identificadas por medio de técnicas de *clustering*.

Además, existen paquetes de análisis estadísticos para la información criminal, los cuales trabajan sobre GIS . Algunos de ellos son: Spatial and Temporal Análisis of Crime, CompStat y CrimeStat [2] [3]. Entre tanto, técnicas como *Concept Space* han sido implementeadas por el Laboratorio de Inteligencia Artificial de la Universidad de Arizona, en Tucson, para extraer relaciones entre la información policial y así detectar posibles bandas o sospechosos.

¿Abstract?

Concept Space se apoyó en el uso de minería de datos y, concretamente, en *Clustering* jerárquico [4]. Cabe resaltar que los recursos ofrecidos por este tipo de técnicas han dado frutos; entre 1985 y 2002 el Gobierno de Estados Unidos detectó 16 miembros clave de grandes organizaciones delictivas [5].

La Explotación de Información (*Data Mining*) [6] es el proceso mediante el cual se extrae conocimiento comprensible y útil —previamente desconocido— desde bases de datos, en diversos formatos y de forma automática. Entonces, la Explotación de Información plantea dos desafíos: trabajar con grandes bases de datos y aplicar técnicas que conviertan, automáticamente, estos datos en conocimiento [7].

Así mismo, *Data mining* es un elemento fundamental para una técnica más amplia cuyo objetivo es el descubrir conocimiento en grandes bases de datos (en inglés, *Knowledge Discovery in Databases* —KDD) [8][9].

El mayor desarrollo del uso de la Explotación de Información en actividades relacionadas con la auditoría de sistemas tiene que ver con la detección de intrusos en redes de telecomunicaciones. Incluso, en la literatura científica se encuentran antecedentes vinculados a la localización de fraudes usando minería de datos [10].

Este texto hace alusión a un caso específico de fraude asociado con las tarjetas de crédito y conocido comúnmente como la <<clonación de tarjetas>>, circunstancia que representa un riesgo para los clientes adscritos a un banco.

Estado del arte

Teoría sobre las transacciones delictivas

Los seres humanos, al tener capacidad cognitiva, desarrollan una serie de conductas que se pueden definir como <<patrones>> dependiendo de ciertas situaciones. A su vez, el momento en el que se comete un crimen no es la excepción; un grupo de psicólogos determinó que existen patrones de comportamiento asociados a factores como la localización, la hora del día y la temperatura. Dicha información, administrada mediante la minería de datos, permite desarrollar un modelo predictivo sobre las situaciones ideales —escenarios— donde podría suceder un crimen. Para el ejemplo citado se establecen tres escenarios que se identifican con los patrones mencionados: robo de bicicleta, robo con arma de fuego y robo de carteras [11].

En consecuencia, el desarrollo de esta herramienta predictiva genera un impacto positivo en la sociedad ya que le permite —a las fuerzas del orden público— tener tiempos de reacción más rápidos y evitar, de esta manera, retrasarse llegando a la escenas del crimen. No obstante, también puede generar un impacto negativo si se prejuzga erróneamente a un ciudadano debido a mala documentación del sistema (falsificación de documento público, por ejemplo) [12]. La revisión manual y técnica de la prevención de fraudes no detecta algunos de los patrones más prevalentes como el uso de una tarjeta de crédito varias veces, en múltiples locaciones (físicas o digitales) y en poco tiempo.

Factibilidad de usar minería en el caso de una institución bancaria colombiana

La idea central de este proyecto es encontrar transacciones fraudulentas las cuales, por lo general, no son la mayoría. Así pues, como método de solución, una buena alternativa es la implementación de la minería de datos; mediante ella pueden identificarse patrones ocultos y no triviales en grandes bases de datos.

Técnicas y algoritmos usados

Técnicas	Definición
1. Discretización	Proceso en la preparación de datos en el cual los valores continuos se vuelven discretos.
2. Normalización	Proceso en la preparación de datos en el cual los valores se parametrizan dentro de un rango definido de 0 a 1.

3. Algoritmo de árboles de decisión (J48)	Modelo de predicción utilizado en el ámbito de la inteligencia artificial y en la minería de datos cuya función es representar y categorizar una serie de condiciones que ocurren de forma sucesiva para la resolución de un problema. El algoritmo J48 de WEKA es una implementación del algoritmo C4.5, uno de los más populares de minería de datos. Además, se trata de un refinamiento del modelo generado con OneR, supone una mejora moderada en las prestaciones y puede conseguir una probabilidad de acierto ligeramente superior a la del anterior clasificador.
4. Regla de asociación *a priori*	Algoritmo que sólo busca reglas entre atributos simbólicos, hecho por lo cual todos los atributos numéricos deberán ser discretizados previamente.

Metodología del proyecto

El Proceso de Descubrimiento del Conocimiento en Base de Datos (DCDB) resulta complejo ya que no sólo incluye la obtención de los modelos o patrones, sino también la evaluación e interpretación de los mismos [8]. El DCDB es definido en [13] como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, compresibles a partir de los datos".

Entre tanto, las principales tareas del proceso de DCDB son, a grandes rasgos, las siguientes: preprocesar los datos, hacer minería de datos, evaluar los resultados y, por último, presentarlos [15][16][17][18]. En la figura 1 se puede observar que el proceso de DCDB está organizado en 5 fases [19]:

Fase 1: recopilación e integración

A partir de la vista minable entregada previamente y tras un análisis entre el experto en el tema y los investigadores, pudo concluirse la necesidad de realizar nuevamente una recopilación e integración sobre estos 1000 datos —reduciendo la cantidad de atributos a trabajar. De esta reducción se hizo una selección de ocho atributos (de los 39 entregados anteriormente). A continuación se presentan los finales:

```
@attribute IDTARJETA numeric
@attribute TIPO_TARJETA {VISA,MASTERCARD,CIRRUS}
@attribute VALORTRANSACCION numeric
@attribute CUPO {LIBRE,MEDIO,COMPLETO,SOBREGIRO}
@attribute TRANSMAYORQUEPROMTRAN {SI,NO}
@attribute FECHATRANSACCION date yyyy-MM-dd
@attribute HORATRANSACCION date HH:mm:ss
@attribute SOSPECHOSO {N, V}
```

El número de registros de la muestra es 1000.

Descripción de atributos:

Idtarjeta	Corresponde al número único de cada tarjeta de crédito.
Tipo_tarjeta	Determina a qué empresa está asociada la tarjeta (Visa, Mastercard o Cirrus).
Valortransaccion	Especifica el monto de la transacción realizada.

CUPO	Especifica la capacidad de endeudamiento de la tarjeta. Por otra parte, se mide con valores categóricos proporcionales a la relación entre el cupo y el saldo disponible para hacer transacciones. Los valores posibles que puede tomar son: "LIBRE", "MEDIO", "COMPLETO" o "SOBREGIRO".
TRANSMAYOR-QUEPROMTRAN	Atributo que valida si la transacción realizada es mayor que el promedio de las transacciones diarias. Se expresa "SÍ" o "NO".
FECHATRAN-SACCION	Fecha de realización de la transacción.
HORATRANSAC-CION	Hora de realización de la transacción.
SOSPECHOSO	Atributo cuya función es determinar el patrón de desconfianza sobre una tarjeta. Se expresa como "V" o "N".

Fase 2: limpieza, selección y transformación (preparación de datos)

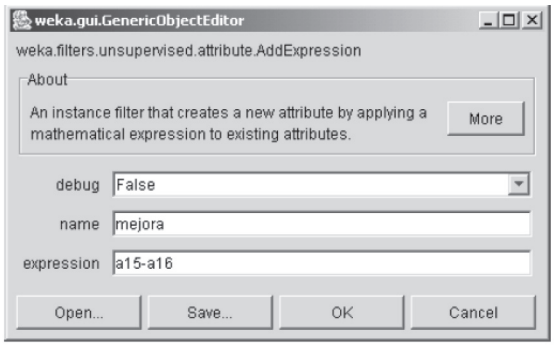
Este caso de estudio fue analizado con una herramienta basada en software libre denominado "Weka".

Mediante el uso de Weka (Waikato Environment for Knowledge Analysis) es posible analizar datos gracias a la aplicación de técnicas de minería de datos. El utilitario Weka contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, los cuales se encuentran unidos a una interfaz gráfica de usuario con fin de acceder fácilmente a sus funcionalidades [20]. Así mismo, este software, que fue desarrollado en la Universidad de Waikato y se puede obtener en forma gratuita en el sitio oficial de esta institución en Internet [22], contiene múltiples algoritmos para la aplicación de técnicas supervisadas y no supervisadas [21][22].

Por otra parte, para la ejecución de la regla de asociación se aplicó, sobre los datos, el proceso de discretización, mientras que para la comparación realizada con el algoritmo de clasificación se realizó el de normalización. Es necesario aclarar que la anterior vista minable no fue nada rápida de sacar, motivo por el cual en la creación de ésta tuvo que realizarse un campo calculado en el archivo arff que tenía el siguiente encabezado:

@attribute TRANSMAYORQUEPROMTRAN {SI,NO}

Para ello se creó un campo adicional llamado <<promedio>> —eliminado más adelante— que contuvo los promedios de las transacciones que realizaría una tarjeta de crédito. Posteriormente, con la ayuda de Weka, fue necesario crear otro campo llamado "TRANSMAYORQUEPROMTRAN" con la opción filtro->no supervisado -> atributo -> adicionar expresión



Este campo valdrá positivo cuando la transacción que se esté llevando a cabo sea mayor que el promedio (valdrá negativo en caso contrario). Una vez efectuado este proceso se reemplazaron los campos que se tenían positivos por el *string* "SÍ" y los campos negativos por el *string* "NO".

En el proceso de preparación de datos también se presentaron algunos problemas. Por ejemplo, la máquina en la que se estaba realizando el proceso de preprocesamiento se colgó y fue necesario cerrar y volver a ejecutar Weka. Este hecho evidenció que Weka posee algunos problemas cuando se manejan grandes cantidades de datos. Por esta razón empezó a trabajarse con una cantidad de datos más pequeña.

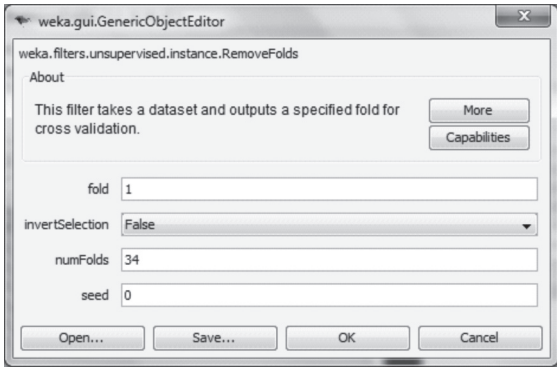
Fase 3: minería específica y aplicación de métodos

Algunos de los métodos existentes:

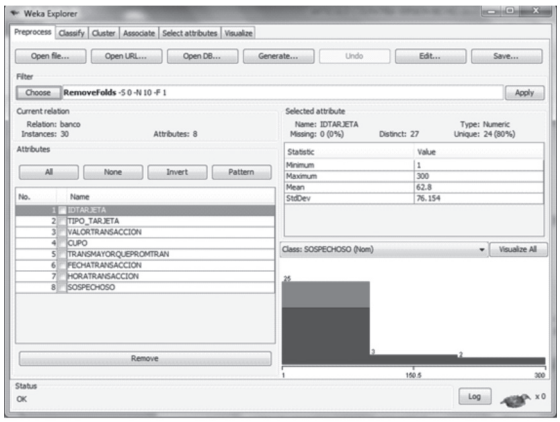
- Modelización estadística.
- Modelización bayesiana.
- Modelos relacionales y declarativos.
- Redes neuronales artificiales.
- Modelos estocásticos y difusos.
- Árboles de Decisión y Sistemas de Aprendizaje de Reglas.
- Modelos basados en núcleo y máquinas de soporte vectorial.
- Modelos basados en casos, densidad o distancia.

Para el caso de detección de fraudes deben elegirse los más relevantes. La clasificación es predictiva y, además, genera un modelo de conocimiento que permite predecir ciertos comportamientos ante la ocurrencia de nuevas situaciones. Dentro de esta técnica, los métodos usados con mayor frecuencia son los árboles de decisión, las redes neuronales y los análisis bayesianos. La asociación de esos tres procedimientos brinda la oportunidad de realizar búsquedas automáticas de reglas que relacionan un conjunto de variables entre sí.

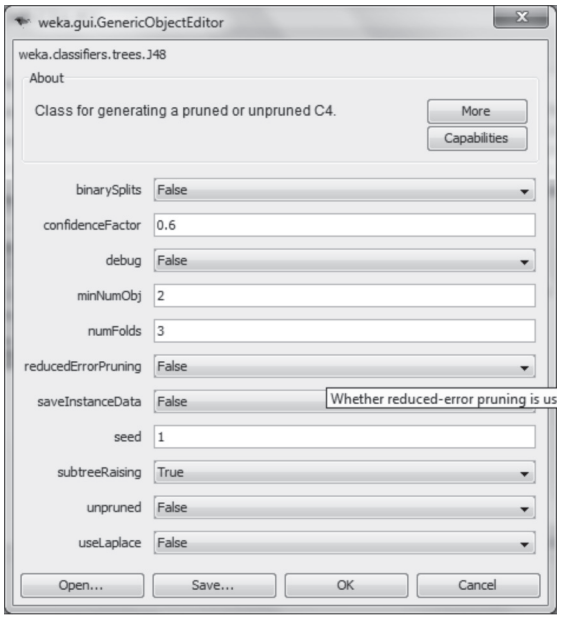
Debido a que el valor de la muestra de datos sospechosos es mucho menor en proporción al resto, se realizó un *Oversampling (*sobre muestreo*)* en la muestra para mejorar la calidad de la regla de asociación. Para ello se utilizó la técnica *RemovedFold* con los siguientes parámetros:



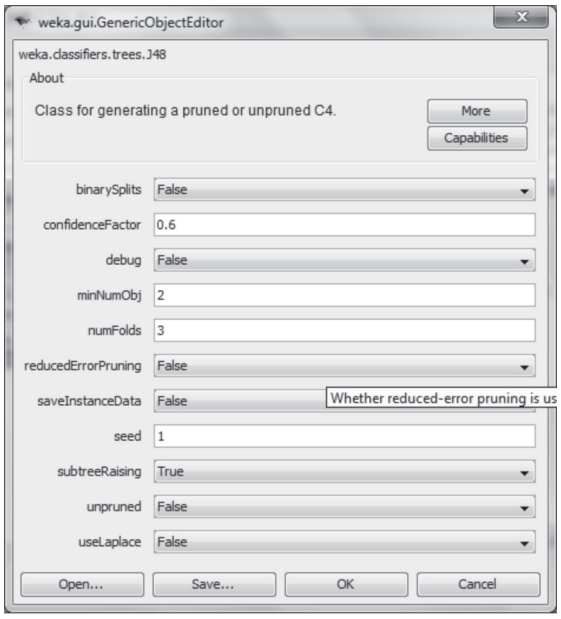
La muestra, entonces, se redujo a 30 instancias tal como muestra la siguiente figura:



Además, se optó por tomar el J48 (implementación en Weka del algoritmo C4.5) [23], así como algunos algoritmos de minería y la Regla de Asociación *a priori*. Para el caso del algoritmo J48 la aplicación se llevó a cabo con los siguientes parámetros:



Por su parte, para la regla de asociación *a priori* se utilizó la misma muestra de 30 instancias. La configuración de los parámetros se realizó así:

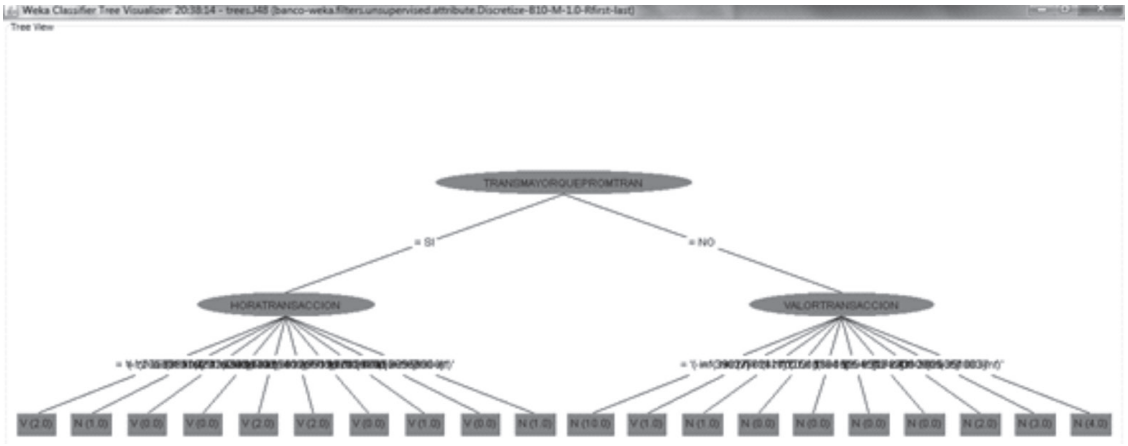


FASE 4: EVALUACIÓN E INTEPRETACIÓN DEL MODELO RESULTANTE

Luego de aplicar el algoritmo de clasificación a los datos se obtuvo la siguiente salida en Weka:

```
=== Run information ===
Test mode:    0-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------
TRANSMAYORQUEPROMTRAN = SI: V (9.0/2.0)
TRANSMAYORQUEPROMTRAN = NO: N (21.0/1.0)

Number of leaves :          2

Size of the tree :   3

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly classified instances          27               90      %
Incorrectly classified instances         3               10      %
Kappa statistic                          0.7541
Mean absolute error                      0.1737
Root mean squared error                  0.3095
Relative absolute error                 43.1657 %
Root relative squared error             69.2638 %
Total Number of Instances               30

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.909    0.125      0.952     0.909     0.93       0.813    N
                 0.875    0.091      0.778     0.875     0.824      0.813    V
Weighted Avg.    0.9      0.116      0.906     0.9       0.902      0.813

=== Confusion Matrix ===

  a  b   <-- classified as
 2 |  a = N
  1  7 |  b = V
```

El algoritmo J48, entre tanto, generó el árbol de decisión que se expone a continuación:



Analizando la matriz de confusión:

```
a  b   <-- classified as
20  2 |  a = N
 1  7 |  b = V
```

Puede observarse que los valores de la diagonal son los aciertos y el resto, los errores. De las 22 tarjetas de crédito no sospechosas, 20 fueron bien clasificadas y sólo 2 no. Por otra parte, de las 8 tarjetas de crédito sospechosas 7 fueron bien clasificadas y 1 fue incorrectamente clasificada.

Regla 1

Si el valor del cupo es cercano a toda su disponibilidad inicial, quiere decir que el cliente no ha hecho uso de la tarjeta de crédito o que pagó la totalidad de su deuda. En cualquiera de los casos se puede concluir que si el usuario no tiene deudas es porque los préstamos o retiros que ha realizado no han estado por encima del cupo de la tarjeta.
Evaluación de la regla: coherente.

Regla 2

Si el cliente tiene casi la totalidad (o entera) del cupo, entonces no posee deudas ni ha realizado mayores transacciones. Con esto la probabilidad de que haya realizado movimientos sospechosos es mínima.
Evaluación de la regla: coherente.

Regla 3

El valor de la transacción puede afectar, en algún grado, la probabilidad de que sea fraudulenta; entre mayor sea el valor del retiro, mayor será la probabilidad del fraude. Sin embargo, no es la única variable que se tiene porque el valor de la transacción está íntimamente ligado al del cupo (para determinar el grado de sospecha de la transacción). Entonces, la regla no es suficiente y es preferible omitirla.
Evaluación de la regla: insuficiente.

Regla 4

Para que esta regla sea totalmente coherente es necesario tener en cuenta el cupo y no el saldo.
Evaluación de la regla: insuficiente.

Regla 5

Evaluación de la regla: insuficiente.

Regla 6

Si el cupo está libre y las transacciones no sobrepasan el promedio usual, es casi seguro que no se trata de un cliente al que le clonaron su tarjeta. Esta es una de las reglas más coherentes.

```
=== Run information ===
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.35 (10 instances)
Minimum metric <confidence>: 0.4
Number of cycles performed: 13

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 11

Size of set of large itemsets L(3): 4

Best rules found:
1. CUPO=LIBRE 10 ==> TRANSMAYORQUEPROMTRAN=NO 10    conf:(1)
2. CUPO=LIBRE 10 ==> SOSPECHOSO=N 10    conf:(1)
3. VALORTRANSACCION='(-inf-39027]' SOSPECHOSO=N 10 ==>
TRANSMAYORQUEPROMTRAN=NO 10    conf:(1)
4. VALORTRANSACCION='(-inf-39027]' TRANSMAYORQUEPROMTRAN=NO 10 ==>
SOSPECHOSO=N 10    conf:(1)
5. CUPO=LIBRE SOSPECHOSO=N 10 ==> TRANSMAYORQUEPROMTRAN=NO 10    conf:(1)
6. CUPO=LIBRE TRANSMAYORQUEPROMTRAN=NO 1 ==> SOSPECHOSO=N 10    conf:(1)
7. CUPO=LIBRE 10 ==> TRANSMAYORQUEPROMTRAN=NO SOSPECHOSO=N 10    conf:(1)
8. HORATRANSACCION='(61279500-69199600]' SOSPECHOSO=N 10 ==>
TRANSMAYORQUEPROMTRAN=NO 10    conf:(1)
9. TRANSMAYORQUEPROMTRAN=NO 21 ==> SOSPECHOSO=N 20    conf:(0.95)
10. SOSPECHOSO=N 22 ==> TRANSMAYORQUEPROMTRAN=NO 20    conf:(0.91)
```

Evaluación de la regla: coherente total.

Regla 7

No es suficiente porque al tener un cupo grande son muchas las posibilidades de que se ejecuten transacciones por encima del promedio.
Evaluación de la regla: insuficiente.

Regla 8

No es una regla muy coherente.
Evaluación de la regla: no coherente.

Regla 9

Evaluación de la regla: coherente total.

Regla 10

Si un cliente no es sospechoso existe una gran probabilidad de que sus próximas transacciones se encuentren dentro del patrón de conducta cotidiano.
Evaluación de la regla: coherente.

Pruebas realizadas y validación

El entrenamiento del árbol se realizó con los 30 datos resultantes luego de aplicar la técnica RemovedFolds —debido a la poca presencia de datos claves para la predicción de fraude. Así mismo, pudo comprobarse la calidad del modelo con un set de pruebas de 1000 registros en donde se presentaron los siguientes resultados:

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.6 -M 2
Relation:     banco-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last
Instances:    30
Attributes:   8
              IDTARJETA
              TIPO_TARJETA
              VALORTRANSACCION
              CUPO
              TRANSMAYORQUEPROMTRAN
              FECHATRANSACCION
              HORATRANSACCION
              SOSPECHOSO
Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

J48 pruned tree
------------------

TRANSMAYORQUEPROMTRAN = SI
HORATRANSACCION = '(-inf-29599100]': V (2.0)
HORATRANSACCION = '(29599100-37519200]': N (1.0)
HORATRANSACCION = '(37519200-45439300]': V (0.0)
HORATRANSACCION = '(45439300-53359400]': V (0.0)
HORATRANSACCION = '(53359400-61279500]': V (2.0)
HORATRANSACCION = '(61279500-69199600]': V (2.0)
HORATRANSACCION = '(69199600-77119700]': V (0.0)
HORATRANSACCION = '(77119700-85039800]': V (1.0)
HORATRANSACCION = '(85039800-92959900]': V (0.0)
HORATRANSACCION = '(92959900-inf)': N (1.0)
TRANSMAYORQUEPROMTRAN = NO
VALORTRANSACCION = '(-inf-39027]': N (10.0)
VALORTRANSACCION = '(39027-78024]': V (1.0)
VALORTRANSACCION = '(78024-117021]': N (1.0)
VALORTRANSACCION = '(117021-156018]': N (0.0)
VALORTRANSACCION = '(156018-195015]': N (0.0)
VALORTRANSACCION = '(195015-234012]': N (0.0)
VALORTRANSACCION = '(234012-273009]': N (0.0)
VALORTRANSACCION = '(273009-312006]': N (2.0)
 VALORTRANSACCION = '(312006-351003]': N (3.0) |    VALORTRANSACCION = '(351003-inf)': N (4.0)

Number of leaves  :           20
Size of the tree :    23

Time taken to build model: 0.01 seconds

=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        1000               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
               1          0          1          1          1          1        N
               1          0          1          1          1          1        V
    Weighted Avg.    1          0          1          1          1          1

=== Confusion Matrix ===
a   b   <-- classified as
789   0 |    a = N
0 211 |    b = V
```

Observando la matriz es notable que el grado de precisión está expresado. La muestra general de 1000 datos manejaba un porcentaje de aproximadamente el 3% de los mismos para contrarrestar los supuestos de no fraude; es decir, de los datos sospechosos.

Estos resultados fueron enseñados al experto del banco, quien estuvo de acuerdo en que la muestra de datos original ofrecía errores al ser generada a partir de dos bases de datos distintas —logrando confundir al experto en la selección de los atributos para generar la predicción. De esta manera, por más preparación de los datos que se hiciera, la calidad de estos no mejoraría.

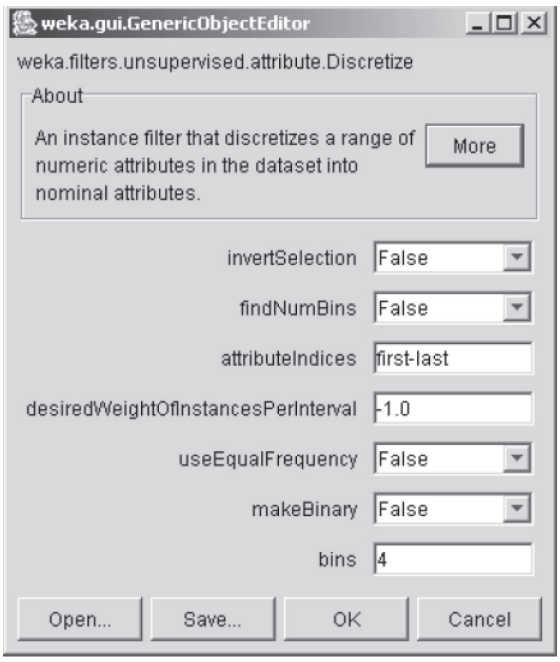
Observando los resultados generados por las reglas de clasificación (algoritmo J48) se esperaría que parte de las reglas de asociación pudieran verse reflejadas dentro del esquema del árbol; sin embargo, dicha situación no se presentó. Entonces, las reglas de asociación fueron menos efectivas con respecto a las de clasificación, hecho que puede notarse en la medida en que se obtuvieron reglas de asociación incoherentes.

Conclusiones y trabajo futuro

La calidad de las reglas de asociación, muchas veces, está delimitada por la presencia de atributos fuertemente descompensados. Para el caso de fraude serían la identificación de la tarjeta y el tipo de la misma, provocano, de esta manera, que no aparezcan en las reglas de asociación (poseen una cobertura baja y son filtradas).

Los filtros de discretización fueron muy útiles cuando se trabajaron con atributos numéricos; varias herramientas de análisis requieren datos simbólicos y, por tanto, se necesita aplicar esta transformación antes. Así mismo, son necesarios cuando debe efectuarse una clasificación sobre un atributo numérico como, por ejemplo, el caso presentado en este proyecto.

En el proceso de discretización no supervisado de atributos pudieron haberse fijado todas las cajas de la misma anchura; no obstante, para la distribución de los datos que se tenía, resultó mucho más útil no hacerlo; es decir, forzar a una distribución uniforme de instancias por categoría. Para ello se empleó la opción de Weka "*useEqualFrequency*" con lo cual pudo determianrse que modificar este parámetro influye notablemente en los resultados obtenidos con los algoritmos.



El número de categorías (*bins*) usadas para efectuar la discretización influyó notablemente en el número de reglas coherentes arrojadas por el algoritmo de asociación y, además, en los resultados obtenidos por los algoritmos de clasificación.

Para fines de este ejercicio se hizo necesario, por cada algoritmo a aplicar, poseer una vista minable. Probablemente, en una situación más compleja en donde se tenga planeado

usar varios algoritmos, deba emplearse un mayor número de vistas minables. Lo anterior eleva la complejidad de los proyectos de minería de datos y genera que dicho proceso esté supeditado, en un porcentaje alto, a la destreza con la que se haga el proceso de preparación de datos.

Por su parte, para el manejo del algoritmo de asociación es estrictamente necesario que los datos a asociar están previamente discretizados, mientras que para la técnica de clasificación deben estén normalizados previamente.

Un estudio que implique seguimiento de transacciones fraudulentas tiene una duración finita que se puede definir en la etapa de planeación. Además, es posible que algunos clientes del banco no hayan realizado transacciones fraudulentas al momento de concluir el estudio. Por otra parte, no es posible determinar el tiempo en el que estos clientes se comportan sin realizar transacciones delictivas; entonces, podría omitirse la información de aquellos que no presentan transacciones delictivas, hecho poco adecuado porque al descartarlos se perdería la información acerca del tiempo durante el cual han estado afiliados al banco —dato de gran utilidad que debe ser incorporado en el análisis. A este tipo de datos se le conoce como <<censuras>> o <<casos censurados>>. Por lo anterior, es sugerible tener en cuenta dichos aspectos en trabajos futuros.

La necesidad de una predicción de fraude sobre unos datos bancarios requiere que la selección de los atributos esté basada en la necesidad de encontrar una variable dependiente respecto a unas independientes; por tanto, se recomienda tener un control específico en la migración de los datos cuando estos no se traen sobre una misma base de datos.

Por consiguiente, es necesario generar consultas sobre la base de datos de manera previa al proceso de preparación de los mismos; sucesos como promedios, diferencias o lapsos son decisivos para estimar patrones de comportamiento en este caso de estudio.

El proceso de minería descrito en este artículo no habría tenido éxito sin la creación del campo TRANSMAYORQUEPROMTRAN; así pues, no está de más suponer que debe contarse con algunas nociones de arte y olfato de minero para saber en qué momento crear campos calculados (gracias a ellos los algoritmos producen los datos requeridos).

Si el banco quisiera adoptar soluciones más robustas para determinar con mayor dinamismo comportamientos sospechosos y fraudulentos por parte de sus clientes o de terceros, tendría que hacer uso de otras herramientas más robustas que Weka como, por ejemplo, soluciones de ETL (extracción, transformación y carga); pueden obtenerse mejores resultados. Entre ellas se destacan:

a. WPS (World Programming System) [25]: puede reemplazar a SAS en empresas con una dependencia importante respecto a este producto. Sus ventajas principales son de compatibilidad de código y, fundamentalmente, de precio.
b. Kettle [26]: sumamente intuitiva y con una curva de aprendizaje prácticamente plana. Su costo es nulo.
c. Talend [27]: parecida a la anterior. Es gratuita y de código abierto que incluye, además, módulos sofisticados de depuración de datos.

Como aspectos interesantes se destaca uno primordialmente: en el momento de extraer la información de una base de datos, el proceso

de limpieza se llevó cabo con la herramienta Weka y, en ningún caso, los datos se consolidaron en un *datawarehouse*; entonces, el tipo de solución propuesto resulta demasiado artesanal y poco escalable. De hecho, al realizar otro proceso de minería (no cuentan con un repositorio único para obtener sus datos) como el descrito, debe extraerse la información nuevamente. Por ende, la implementación de un *datawarehouse* como repositorio en donde residan los datos históricos le brindaría a este proyecto una mayor escalabilidad a futuro.

Agradecimientos

A todas las personas que han contribuido en mi formación bancaria y a la entidad que me facilitó los datos (por razones de confidencialidad no puede mencionarse el área ni la institución). Sin esta información no habría sido posible desarrollar este artículo en el que se demuestra cómo, mediante la minería, pueden llevarse a cabo proyectos tan reales y álgidos como la seguridad informática.

Referencias

Eck, J. *et. al.* (2005), "Mapping Crime: Understanding Hot Spots. Crime Mapping Research Center U.S. Department of Justice. Office of Justice Programs".

ICJIA, (2007), "Illinois Criminal Justice Information Authority", [en línea], disponible en: http://www.icjia.state.il.us/public/index.cfm?metasection=Data&metapage=StacFacts, recuperado el 1 de marzo de 2011.

CrimeStat (2007), "CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations", [en línea], disponible en: http://www.icpsr.umich.edu/NACJD/crimestat.html, recuperado el 1 de marzo de 2011.

Coplink (2007), "COPLINK Solution Suite", [en línea], disponible en: http://www.coplink.com, recuperado el 1 de marzo de 2007.

Coplink, (2004), Crime Data Mining and Visualization for Intelligence and Security.

Informatics: The COPLINK Research. University of Arizona Artificial Intelligence Lab. URL: http://ai.bpa.arizona.edu/research/coplink/index.htm. Acceso Marzo 2011.

Clark, P. y Boswell R. (2000), *Practical Machine Learning Tools and Techniques with Java Implementation*, CIUDAD, Morgan Kaufmann Publisher.

Britos, P.; Dieste, O. y García, R. (2008), "Requirements Elicitation in Data Mining for Business Intelligence Projects" en *Advances in Information Systems Research, Education and Practice*, VOL., NÚM, pp. 139 – 150.

Fayyad U.M.; Piatetsky, G. y Smyth, P. (1996), "From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining" NOMBRE, VOL, NÚM, pp. 1-34.

Britos, P. *et al.* (2005), *Minería de Datos Basada en Sistemas Inteligentes*, CIUDAD, Nueva Librería.

Britos, P. *et al.* (2008), Detecting Unusual Changes of Users Consumption. In Artificial Intelligence and Practice II. Springer. p. 297-306.R. W. Lucky, "Automatic equalization for digital communication," *Bell Syst. Tech. J.*, vol. 44, no. 4, pp. 547–588, Apr. 1965.

Gunderson, L. (2002), "Using data mining and judgment analysis to construct a predictive model of crime" IEEE International Conference, CIUDAD.

Brown, D. y Oxford, R. (2001), "Data mining time series with applications to crime analysis," IEEE International Conference, CIUDAD.

Hernandez, O.; Ramírez, Q. y Ferri, R. (2004), *Introducción a la Minería de Datos*, Madrid, Editorial Pearson Prentice Hall.

Fayyad, U.; Piatetsky-Shapiro G. y Smyth P. (1996), *From Data Mining to Knowledge Discovery: An Overview*. *Advances in Knowledge Discovery and Data Mining*, CIUDAD, AAAI/MIT Press.

Chen, M.; Han J. y Yu P. (1996), "Data Mining: An Overview from Database Perspective", en *IEEE Transactions on Knowledge and Data Engineering*, VOL, NÚM, PP.

Han, J. y Kamber, M. (2006), *Data Mining Concepts and Techniques*, San Francisco, Morgan Kaufmann Publishers.

Hand, D.; Mannila, H. y Smyth P. (2001), *Principles of Data Mining*, CIUDAD, MIT Press.

Timarán, P. (2009): "Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos", Memorias de la VIII Conferencia Iberoamericana en Sistemas, Orlando, Estados Unidos de América.

Hernandez, O.; Ramí rez, Q. M. y Ferri, R. C. (2004), *Introducción a la Minería de Datos*, Madrid, Editorial Pearson Prentice Hall.

Sierra, B. et al. (2006), *Aprendizaje automático: conceptos básicos y avanzados: aspectos prácticos utilizando el Software WEKA*, CIUDAD, Pearson.

Dapozo, G. (2006), "Aplicación de minería de datos con una herramienta de software libre en la evaluación del rendimiento académico de los alumnos de la carrera de Sistemas de la FACENA-UNNE", Anales del Octavo Workshop de Investigadores en Ciencias de la Computación WICC 2006, Buenos Aires, Argentina.

Machine Learning Project at the Department of Computer Science of the University of Waikato, New Zealand. http://www.cs.waikato.ac.nz/ml/weka/

Witten, I. y Frank, E. (2005), *Data Mining Practical Machine Learning Tools and Techniques*. San Francisco, Morgan Kaufmann Publishers.

http://www.teamwpc.co.uk/products/wps

http://kettle.pentaho.com/

http://www.talend.com/