The Integrated Data Mining Tool MineKit and a Case Study of its Application on Video Shop Data.

Joel Larocca Neto   Alexandre D. Santos
Celso A.A. Kaestner   Alex A. Freitas

Pontificia Universidade Catolica do Parana
Programa de Pos-Graduacao em Informatica Aplicada
Rua Imaculada Conceicao, 1155
Curitiba - PR. 80215-901. Brazil.
{joel, denes, kaestner, alex}@ppgia.pucpr.br
http://www.ppgia.pucpr.br/~alex

1. Introduction

In essence, data mining consists of extracting high-level, *interesting* knowledge from real-world data sets. It is a very interdisciplinary field, combining ideas and methods of several areas, such as machine learning, statistics and databases.

Data mining is the core phase of a broader process, called the knowledge discovery process [Fayyad et al. 1996]. This process also includes data preprocessing (data cleaning, attribute selection, etc.) and discovered-knowledge postprocessing activities. It is estimated that data preprocessing activities take about 80% of the time dedicated to the entire knowledge-discovery process.

This paper has two main goals. The first one is to introduce MineKit, an integrated data mining tool. In the spirit of the interdisciplinarity of data mining, MineKit contains several algorithms drawn from different research areas and offers a nice 3-D visualization interface, as will be seen later.

The second goal of this paper is to report the result of evaluating MineKit in a real-world data set. This case study is relevant for data mining mainly for two reasons. First, the original data set, like a typical real-world data set, was not previously prepared for data mining activities, so that we had to spent a significant time preparing the data. Hence, we have actually gone through the most time-consuming phase of the knowledge discovery process. This issue is usually ignored in the data mining literature, which focus on the data mining phase only.

The second reason why our case study is relevant is that we have discovered knowledge that is not only accurate and comprehensible, but also interesting, in the sense of being novel and surprising for the user. This is another issue which is often ignored in the literature, which tends to focus on the predictive accuracy and comprehensibility of the discovered knowledge. To see that predictive accuracy and comprehensibility do not imply interestingness, consider a hospital database. It is quite possible that a data mining algorithm discover in this kind of database the following rule: IF (patient_is_pregnant? = “yes”) THEN (patient_gender = “female”). This rule is highly accurate and highly comprehensible, but is completely uninteresting for the user, since it states an obvious, previously-known relationship. An additional discussion on rule interestingness can be found e.g. in [Freitas 1999].

The rest of this extended abstract is organized as follows. Section 2 presents an overview of MineKit. Section 3 describes the data set used in our case study and some preprocessing applied to that data set, to prepare it for data mining. Section 4 briefly reports our computational results. Section 5 concludes the paper.

2. An Overview of MineKit

MineKit is an integrated tool combining standard data mining algorithms from several research areas. Briefly, the algorithms incorporated in MineKit are as follows:

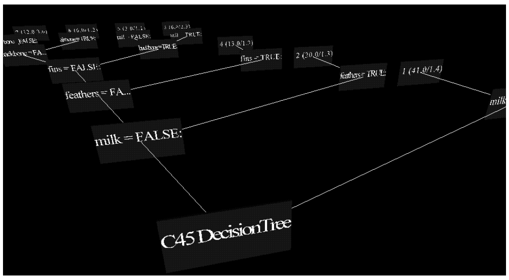
* An algorithm for discovering association rules, based on the Apriori algorithm [Agrawal et al. 1996].

* Three algorithms for classification, more precisely: (a) the C4.5 decision-tree algorithm [Quinlan 1993]; (b) a multi-layer perceptron / backpropagation neural network algorithm [Rumelhart & McClelland 1986], which can be run with any number of neurons and hidden layers, as specified by the user; and (c) a k-nearest neighbor (or instance-based learning) algorithm with weighted attributes [Dasarathy 1991], where the attribute weights are set by the user.

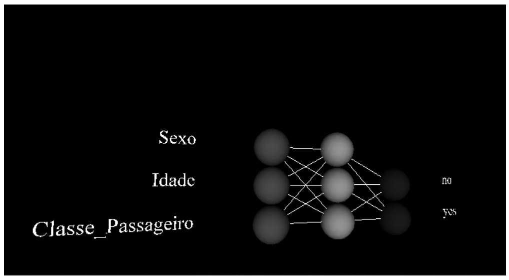
* A clustering algorithm, based on the well-known k-means algorithm [Witten & Frank 2000].

It should be noted that MineKit's algorithms not only cover a range of different different data mining tasks (association, classification and clustering), but also are representative instances of several knowledge discovery paradigms. This is a crucial point, since it has been shown - both theoretically [Schaffer 1994] and empirically [Michie et al. 1994] - that no single algorithm or knowledge discovery paradigm is "the best" across all applications domains. In other words, the effectiveness of a data mining algorithm strongly depends on the data set being mined.

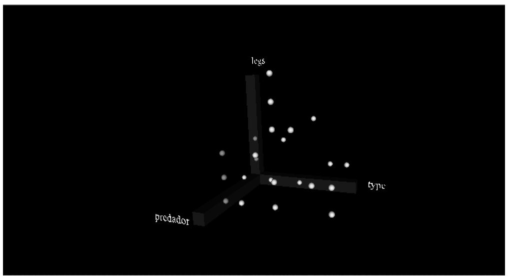
Therefore, in practice it makes sense to offer the user a collection of different algorithms, so that the user can apply several algorithms to his/her data and simply pick up the best result among all the results produced by the individual algorithms. This multi-algorithm approach is precisely the approach followed by MineKit.



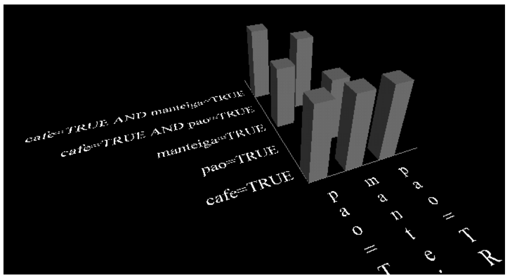
(a) decision trees



(b) neural nets



(c) k-means clustering



(d) association rules

Figure 1: 3-D visualization of discovered knowledge in MineKit

In addition to the above-mentioned five kinds of algorithm, MineKit also contains a basic set of classes which allows the user to develop new data mining algorithms and add them to the tool. Furthermore, MineKit also has nice 3-D visualization resources, developed in VRML (Virtual Reality Modeling Language). A sample of the visualization resources offered by MineKit is shown in Figure 1. Finally, MineKit offers several facilities for interfacing with industrial-strength relational database systems, such as: it is compliant with the ODBC (Open Database Connectivity) standard; it allows de-normalization of relations via a foreign key; it allows insertion, removal and update of tuples; and it performs data filtering.

3. Data Preprocessing for Data Mining

The data set used in our case study contains data about customers and videotape rentals in a video shop. The data was originally stored in third normal form, and it was spread through ten tables of a relational database system. Out of these ten tables, four were effectivily used for data mining purposes. Since data mining algorithms usually require de-normalized data concentrated in a single table, we have used MineKit to put the data in first normal form.

We have opted for creating two first-normal-form tables, one of the them containing data about individual videotape rentals and the other one containing data about customers. Then we can apply data algorithms to each of these tables separately, as will be seen later. The attributes of these tables are described in Table 1 and Table 2.

Table 1: Attributes of the videotape rentals table

Attribute	data type and domain values
week day	categorical (1 ... 7)
month	categorical (1 ... 12)
age	continuous or categorical (0..30; 31..45; 46..60; 60...)
district	categorical (155 different categories of districts)
movie category	categorical (15 different categories of movies)

Table 2: Attributes of the customer table

Attribute	data type and domain values
district	categorical (155 different categories of districts)
age	continuous or categorical (0..30; 31..45; 46..60; 60...)
total number of videotape rentals	categorical (0...100; 101...200; 201...300; 301...500; 500...)

The first two attributes of Table 1 refer to the day of the week and the month of a given videotape rental. These attributes were treated as categorical (or nominal). The third attribute, the age of the customer who rented the corresponding videotape, was treated either as a continuous or as a categorical attribute, depending on the data mining algorithm being used. In the latter case the categorical values consist of age intervals (as shown in the table) which were empirically determined. The fourth attribute is the name of the district of the city where the customer who rented the videotape lives. The fifth attribute is the movie category of the rented videotape.

The first two attributes of Table 2 indicate respectively the district where a customer lives and his/her age. The third attribute is the total number of videotapes rented by the customer, since he/she became a registered customer of the videoshop.

Note that the customer table has one row per customer, while the videotape rental table has one row per videotape rental. This latter table contains data about videotapes rented in the last six months. These tables have 5075 and 3062 rows respectively.

In addition to creating the data mining-oriented tables of videotape rental and customer (whose structure is shown in Tables 1 and 2), we have also had to do a lot of data cleaning. The original, normalized tables contained a large number of attributes, most of which had many incorrect or missing values. Also, many attributes (such as videotape-id, customer's document numbers, etc.) were irrelevant for data mining, and so they were not used used to produce the data mining-oriented videotape rental and customer tables. Some attributes which had to be treated as categorical attributes, such as district, contained free textual values, leading to an excessive number of distinct values. We have had to manually correct these attribute values. We also had to discretize some attributes, such as age, since some data mining algorithms (e.g. the algorithm for discovering association rules) cannot directly cope with continuous attribute values.

4. Computational Results

We have performed 4 experiments, applying different MineKit algorithms to the videotape rental and customer tables described in the previous section.

Experiment 1 - We used C4.5 to produce a decision tree that predicts the category of a movie to be rented by a customer, mining the data of the videotape rental table. The overall accuracy rate of the decision tree on the test set was 68%. Some interesting rules discovered by C4.5 are as follows:

IF Age_Range = up_to_30 THEN Movie_Category = "Comedy"

IF Age_Range = from_31_to_45 THEN Movie_Category = "Childish"

IF Age_Range = from_46_to_60 THEN Movie_Category = "Comedy"

These rules shown that: (a) comedy is a movie category popular across distinct Age ranges; and (b) childish movies are often rented by middle-aged people, who probably have young children.

Experiment 2 - We used the Apriori algorithm to discover association rules with at least 80% of confidence and 2% of support, mining the data of the videotape rental table. Some interesting discovered rules are as follows:

IF District = "Alto_XV" THEN Age_Range = from_46_to_60 (confidence = 86%)

IF District = "Sao_Lourenco" THEN Age_Range = from_31_to_45 (conf. = 100%)

IF Movie_Category = “Childish” THEN Age_Range = from_31_to_45 (conf. = 80%).

The first two rules show that, for the two corresponding districts, most customers belong to a specific age range. The third rule is strongly related to the second rule discovered by C4.5, reported above in Experiment 1. The discovery of these two rules, complementary to each other, reinforces our belief in the correlation between childish movies and the age range from 31 to 45 years old.

Experiment 3 - We used C4.5 to produce a decision tree that predicts the range of total number of videotapes rented by the customer, since he/she became a registered customer of the videoshop, mining the customer table. The overall accuracy rate of the decision tree on the test set was 81%. Two interesting rules discovered by C4.5 are as follows:
IF District in (jard_das_americas, c_imbuia, p_velho) THEN class = more_than_500
IF District in (jd_campo_alto_atu, j_sao_paulo, vila_centenario, champagnat, prive_bois_de_bolo, edif_morada_s_diego) THEN class = from_301_to_500

The names of districts occuring in the above rules probably do not make sense for the reader. However, they do make sense for the owner of the videoshop. Actually the above two rules can be considered the most interesting piece of knowledge discovered in our case study, in the sense of representing knowledge that is highly surprising (or novel) for the user. The reason is that the district names occurring in the above rules represent areas of the city which are relatively far away from the location of the videoshop. In other words, the most profitable customers of the videoshop do not live nearby, they come from relatively distant and well-defined districts.

Experiment 4 - We have used a single-layer perceptron neural net (i.e. with no hidden layer) to predict the range of total number of videotapes rented by the customer, mining the customer table. This is the same task as addressed in experiment 3. The neural net achieved an accuracy rate of 80% on the test set. This is very similar to the performance of C4.5 in experiment 3. However, C4.5 discovers high-level, comprehensible rules, whereas the output of the neural net represents low-level knowledge, which cannot be directly interpreted by the user.

5. Conclusions.

We have introduced MineKit, an integrated data mining tool containing algorithms from several research areas, such as machine learning, statistics and 3-D visualization. We have used MineKit to mine a real-world database containing data about videotape rentals and customers of a video shop.

Some of the rules discovered by MineKit were truly interesting, in the sense that they represent knowledge that was surprising and novel for the user. We consider this an important result, since the discovery of truly interesting knowledge is usually ignored in the literature, which tends to focus on predictive accuracy and rule comprehensibility - as discussed in the Introduction.

References.

Agrawal R, Mannila H, Srikant R, Toivonen H and Verkamo AI. Fast discovery of association rules. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R. (Eds.) *Advances in Knowledge Discovery and Data Mining*, 307-328. AAAI/MIT Press. 1996.

Dasarathy BV. (Ed.) *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* IEEE Computer Society Press, 1991.

Fayyad UM, Piatetsky-Shapiro G and Smyth P. From data mining to knowledge discovery: an overview. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R. (Eds.) *Advances in Knowledge Discovery and Data Mining*, 1-34. AAAI/MIT Press. 1996.

Freitas AA. On rule interestingness measures. *Knowledge-Based Systems 12(5-6)*, 309-315. Oct. 1999.

Michie D, Spiegelhalter DJ and Taylor CC. *Machine Learning, Neural and Statistical Classification*. New York: Ellis Horwood, 1994.

Quinlan JR. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

Rumelhart DE and McClelland JL. (Eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Cambridge, MA: MIT Press, 1986.

Schaffer C. A conservation law for generalization performance. *Proc. 11th Int. Conf. Machine Learning*, 259-265. 1994.

Witten IH and Frank E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, 2000.