Automation of Financial Loaning System
An Analysis of Classification Algorithms on Different Data Views

SAIMA MUSHTAQ, SOBAN HAMEED, AMNA ASAD
LIAQUAT MAJEED SHEIKH
Center of Knowledge Engineering
National University of Computer and Emerging Sciences
Lahore-PAKISTAN

Abstract: - Supervised learning plays a significant role in predicting the behavior of new data, based on the rules, which are extracted keeping in view the behavior of existing data in the database. This paper is about algorithmic analysis of supervised learning for transactional data. Our main idea is to apply different classification algorithms on a preprocessed financial data set in order to evaluate that which type of classification algorithm under what sort of data model selection and with what combination of mining attributes is best suited for a transactional and frequently occurring data. In this way the algorithm with highest accuracy can be used to predict the credit rating of a client, based on his past transactions. It can be very helpful for a financial institute to develop an automated loaning system with least chance of error and fraud.

Keywords: - Data-View, ID3, C4.5, Accuracy, Financial Data

1 Introduction

Normally decisions for giving loans and other services to specific clients are carried out by the staff of a financial institution manually with great chances of error and fraud. These chances can be reduced by the automation of the prevalent loaning system of financial institutions, through which banks can rate clients for a specific favor. The prerequisite of doing this task is the availability of information about the clients, their accounts, their monthly and daily transactions, granted loans and issued credit cards etc. The banks generally save information about their clients and accounts. This stored information can be mined to check the existing *patterns* of data. These patterns can be very useful in taking decision whether to facilitate a specific client based on his past transactional behavior and defining new rules and strategies for loan services offered by a financial institution.

In this paper we have analyzed a transactional dataset for the preset loaning system by mining the stored data. The first step was preprocessing of data, which included cleaning, pruning and normalization of the data. During this process some new composite features were extracted from the data, which were used in identification of patterns. After preprocessing step we have added on step to the process, which is that the data was viewed logically in consideration with the personal, social, financial, psychological and working status of a client of the financial institution. Our claim is that accuracy of prediction can be improved in this way. On this preprocessed data supervised learning algorithms were applied to know the algorithm with highest accuracy. The algorithms used were ZeroR, ID3 and C 4.5 and these are implemented in WEKA 3.4, which is a collection of machine learning algorithms for data mining tasks.

The rest of the paper is organized as follows: Section 2 & 3 give introduction of classification and classification algorithms. Section 4 tells us about the source, nature, description and preprocessing of the dataset which was used for mining. In section 5 the process of knowledge discovery is discussed involving description of the evaluation methodology used and formation of, different logical views of the dataset. Section 6 reviews the results of classification algorithms.

2 Classification

Classification is one way of supervised learning and is the most commonly applied data mining technique, employs a set of preclassified examples to develop a model that can classify the population of records at large. This type of analysis is very productive in developing strategies for Fraud prevention and credit-risk, where we have heavy transactions by clients most of the time. For a fraud detection application, this would include complete records on a record-by-record basis. The classification algorithm uses these preclassified transactions to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. For example, a classifier capable of identifying risky loans could be used to aid in the decision of whether to grant a loan to a client. [3]

3 Classification Algorithms

In our paper we have compared classification algorithms with the perspective of accuracy measurement. Among these algorithms ZeroR is a simple classifier. But ID3 and C4.5 are based on top-down recursive divide-and-conquer technique, where a decision tree is built. At start, all the training records are at the root. These records are partitioned recursively based on the test attributes. Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain).

ZeroR (Zero Rule) is the simplest classification method. It finds the classifier with maximum frequency in the training data and assigns this class to all the existing records. All future data are classified according to that most frequent classifier. It is a less sophisticated classification algorithm; however it provides a good baseline to compare against. By comparing the accuracy of (much more sophisticated) classifier versus the ZeroR baseline, we can measure the relative improvement. Hence it can act as an efficiency threshold. Of course, if a classifier performs worse than ZeroR, it might be a matter of concern. [4]

ID3 ID3 is a simple decision tree-learning algorithm developed by Ross Quinlan (1983). The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given set, we introduce a metric which is known as "information gain" To find an optimal way to classify a learning set, what we need to do is to minimize the depth of the tree.

C4.5 Quinlan C4.5 is an extension of the base algorithm ID3. C4.5 builds a decision tree using the standard top-down induction of decision trees approach, recursively partitioning the data into smaller subsets, based on the value of an attribute. At each step in the construction of the decision tree, C4.5 selects the attribute that maximizes the information gain ratio. The induced decision tree is pruned using pessimistic error estimation.C4.5 incorporates numerical (continuous) attributes, nominal (discrete) values of a single attribute may be grouped together, to support more complex tests, supports post-pruning after induction of trees, e.g. based on test sets, in order to increase accuracy and deals with incomplete information (missing attribute values). [2]

4 The Knowledge Discovery Data

4.1 Data Source

The dataset selected for the experimentation was from PKDD'99 Discovery Challenge. PKDD'99 Discovery

Challenge was part of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'99), September 15 - 18, 1999, Prague, Czech Republic.

4.2 Data Description

The selected dataset was transactional history of a financial institution and represented complete schema of a bank [5]. The Financial Data Set consists of 8 relations describing bank accounts, clients, credit cards, permanent orders, transactions, and loans. [5]. Each table consists of several variables characterizing a property that can be associated with an account. Hence the object of the investigation is the account with its properties. The account has static and dynamic characteristic.

4.3 Data Preprocessing

Preprocessing of dataset plays a vital role in discovering some useful knowledge about it. In order to prepare this cumbersome data for the data mining an extensive phase of data preprocessing was carried out, so that only the relevant data could be brought in the mining process. In our financial dataset we were able to identify, by carefully examining, some dimensions that were not relevant to the decision making of Loaning System. Following steps were involved in the process.

a) Date Reduction It is the process of reducing the dataset by bringing into mining process only the relevant data. The following attributes (accompanied with their relation names) from the original relations were eliminated because of their irrelevancy with Decision factor involved in the loaning system.

b) Attribute Construction New attributes are constructed and added from the given set of attributes to help the mining process. Filters were programmed in visual C++ for the accomplishment of above-mentioned task. Some already built WEKA3.4 filters were also used too. Following attributes were introduced in order to have some useful results.

c) Data Aggregation The process of summing up data from multiple tables in such a way that it can provide some useful baseline for mining purposes. In the original data, the table *Transaction* consisted of 1056320 objects and each record describes one transaction on one account [5], which was not helpful in its raw form. So there was a need to sum up the table. Therefore extensive queries were applied in order to get a summarized history of every client. This single table consisted of following aggregated values.

Transaction Summary Attributes
Maximum amount withdrawn for an account.
Minimum amount withdrawn for an account.
Count of *credit payments* for an account
Count of *withdrawn payments* for an account
Maximum balance in an account
Minimum Balance in an account
Count of *household transactions* for an account.
Count of *loan installments* for an account
Count of *sanction interest transactions* for an account
Count of *pension transactions* for each Account
Count of *interest credit transactions* for an account

Table 1 Aggregated data Attributes

The new table was joined with other relational tables by Access database queries, in order to create a final client information table, where each row represented an aggregated and summarized record of every loaned client along with the corresponding account and client identity.

5 The Knowledge Discovery Process

After preprocessing step, we applied the above mentioned classification algorithms on the data. The tool used was WEKA 3.4. WEKA is a collection of machine learning algorithms for data mining tasks. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well suited for developing new machine learning schemes. WEKA is open source software issued under the GNU General Public License.

5.1 Evaluation Methodology

There are two types of models, which are used to evaluate a classification algorithm:
a) The first way of evaluation is to split the given data into testing and training data. A part of the data (training set) is labeled and then classification rules are extracted.
b) *Cross fold validation* It is a dynamic technique used to optimize the parameters or features chosen in a classifier. In m-fold cross-validation, the training set is randomly divided into m disjoint subsets with roughly equal size. Each of these n subsets is left out in turn for evaluation, and the other (m - 1) subsets are used as inputs to the classification algorithm. [8] We used 5 fold cross validation for our data.

5.2 Logical Views of Data

We have introduced a new concept of testing different combination of mining attributes of the final client info table in order to predict about different categories of clients. In this way we can predict with more accuracy, because we have classified the data with different logical subgroups. We have called these combinations *views* of the data. When we studied our final client info table then there were five types of factors, which could affect the credit rating of a person if combined in different ways. a) Personal information e.g. age, sex etc. b) Employment status c) Loaned and not loaned clients. d) Client's transactional history. e) Social and psychological factors e.g. crime rate and average salary of a certain place. Our experiments involved six such views, which were chosen keeping in view different factors. The selected views were created is shown in Annex A

5.2.1 Why Data Views?

The data views are selected by the decision making party. In this way we can provide user flexibility, which is an essential requirement of a financial loaning system.
Secondly the idea of using a logical mix of attributes and most relevant attributes to mining process is due to the fact that boosts the classification accuracy by including the desired attributes to act as mining attributes and excluding the undesired attributed from the classification procedure. The results given in the next section proves this statement with respect to different data views.

6 Results and Discussion

By using multiple combinations and applying the three algorithms on every logical view we were able to find out the accuracy differences among them and the algorithm's behavior towards a specific input. We have found that in some cases, if the very relevant (logically inter-related) attributes are brought for the mining process precision may rise to more than 90 percent.

ZeroR As we have discussed earlier that ZeroR is a simple algorithm which counts the frequencies of classes but it can act as baseline for the performance of any other classification algorithm. So if we study the graph of the ZeroR for all the logical views then we can see that in all cases, the lower bound defined by ZeroR for accuracy measurement is 88%.

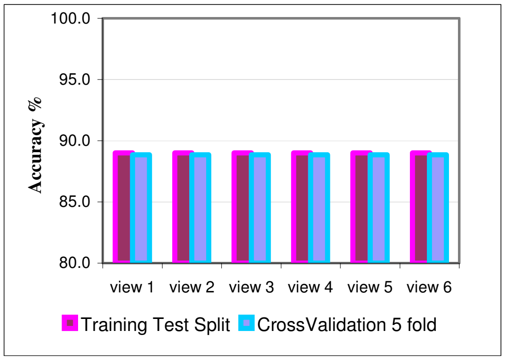

Fig 1.ZeroR Accuracy for different Views

ID3 ID3 initially seemed most relevant to our data, as most of the attributes were in nominal form. However the results given by the algorithm were varying on various inputs as shown in Fig.2. Moreover it did not fulfill the baseline condition defined by ZeroR i.e. its accuracy was most of the time less than 88%. Hence this argument made ID3 least appropriate for our transactional data.

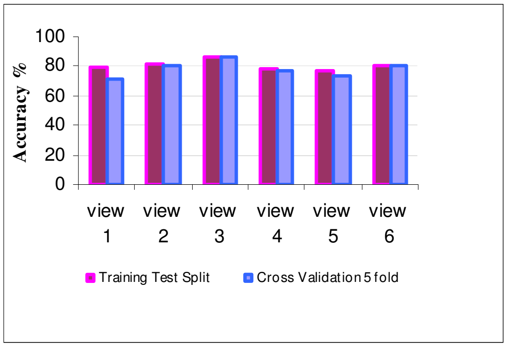

Fig 2.ID3 Accuracy for different Views

C4.5 In WEKA 3.4 C4.5 is implemented with the name J48.When we applied C4.5 then we found that it fulfilled the baseline condition as well as shown consistent behavior with the respect to accuracy. It is clear from Fig. 3 that that by using any logical view as input, C4.5 results in consistent accuracy levels. The results shown in fig also describe that a logical mix of attributes sometimes yields better and consistent classifier performance and in the end increased accuracy. Here pruning of unnecessary nodes of the decision tree may increase the accuracy.

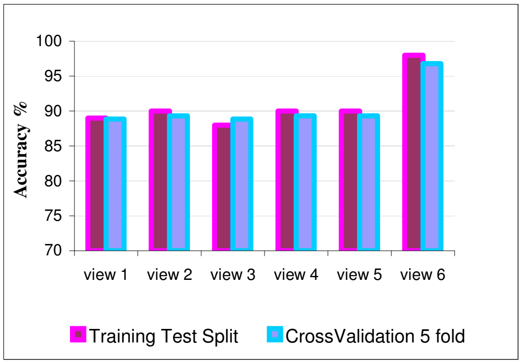

Fig 3. C4.5 Accuracy for different Views

7. Conclusion

In this paper automation of loan decision of a financial institution was focused. We have selected data from financial domain and after necessary preprocessing we have used this data for extracting rules, which are helpful in predicting good or bad credit rating of existing clients who have taken the loan. Now on basis of these rules we have tried to predict the behavior of clients applying for loan in future. Further we have analyzed our data in different logical views which represent combination of personal, social, psychological, financial and employment status of a client in order to view all dimensions of data, which according to us is better instead of getting rules on the whole. By applying the Classification algorithms, rules were extracted from the logical views of data. Results have shown that C4.5 is best algorithm for transactional and frequently occurring data. Our future work includes studying the same patterns through clustering and association algorithms and their comparisons.

References:

[1] J. R. Quinlan. “Induction of decision trees” Machine Learning, 1:81-106, 1986.

[2] J.R. QUINLAN, "C4.5: Programs for Machine Learning", Morgan Kaufmann, 1993.

[3] http://www.dbmsmag.com/9608d53.html

[4] www.chemeng.utoronto.ca/~datamining/Presentations/ ZeroR-OneR.**ppt**)

[5] 3rdEuropean Conference on Principles and Practice of Knowledge Discovery in Databases, September 15 - 18, 1999, Prague, Czech Republic http://lisp.vse.cz/pkdd99/

[6] Michael Spenke, Christian Beilken, “Visual, Interactive Data Mining with InfoZoom – the Financial Data Set”, Proc. 3rdEuropean Conference on Principles and Practice of Knowledge Discovery in Databases , September 15 - 18, 1999, Prague, Czech Republic

[7] David Coufal, Martin Holena, Anna Sochorova, “Coping with Discovery Challenge by GUHA”, Proc. 3rdEuropean Conference on Principles and Practice of Knowledge Discovery in Databases , September 15 - 18, 1999, Prague, Czech Republic

[8] Ka Yee Yeung, Roger E Bumgarner, “Multiclass classification of microarray data with repeated measurements: application to cancer”

Annex A

Views	Description of View	Factors Combined
View1	Credit rating of a male/female client of certain age, with a certain salary range working in a company of certain category, also keeping in view, maximum credited amount, maximum amount withdrawn, his already taken loan amount if any, loan duration and average salary and crime rate in the area where he lives.	Transactional history excluded
View 2	Credit rating of a client of certain age ignoring his/her employee status and keeping in view his/her frequency of household transactions, maximum amount withdrawal, maximum credited amount and sanctioned interest amount. (For not employee people).	Personal and transactional history
View 3	Credit rating of a male/female client of certain age, salary range working in a company of a specific category.	Personal & employment status
View 4	Credit rating of a client of certain age, with a certain salary range working in a company, also keeping in view his transaction frequency, maximum credited amount, maximum amount withdrawn, and sanctioned interest amount	Personal, working status and transactional history included
View 5	Credit rating of a client of certain age, with a certain salary range working in a company, also keeping in view his already taken loan amount if any, loan duration, paid loan installments, interest sanctioned and average salary and crime rate in the area where he lives.	Personal, working status, loan status and social status included.
View 6	Credit rating of a male/female client of certain age, with a certain salary range working in a company of certain category, also keeping in view his transaction frequency, maximum credited amount, maximum amount withdrawn, his already taken loan amount if any, loan duration, paid loan installments, interest sanctioned and average salary and crime rate in the area where he lives.	All five factors included

Table 2 Logical Views of Data