A Case Study of Predicting Banking Customers Behaviour by Using Data Mining

Xujuan Zhou  
School of Management and Enterprise  
University of Southern Queensland  
Springfield, Australia  
xujuan.zhou@usq.edu.au

Ghazal Barghady  
School of Management and Enterprise  
University of Southern Queensland  
Springfield, Australia  
ghazal.barghady@usq.edu.au

Moloud Abdar  
Département d’informatique  
Université du Québec à Montréal  
Montréal, Canada  
m.abdar1987@gmail.com

Xiaohui Tao  
Faculty of Health, Engineering and Sciences  
University of Southern Queensland  
Springfield, Australia  
xtao@usq.edu.au

Raj Gururajan  
School of Management and Enterprise  
University of Southern Queensland  
Springfield, Australia  
raj.gururajan@usq.edu.au

KC Chan  
School of Management and Enterprise  
University of Southern Queensland  
Springfield, Australia  
kc.chan@usq.edu.au

Abstract— Data Mining (DM) is a technique that examines information stored in large database or data warehouse and find the patterns or trends in the data that are not yet known or suspected. DM techniques have been applied to a variety of different domains including Customer Relationship Management (CRM). In this research, a new Customer Knowledge Management (CKM) framework based on data mining is proposed. The proposed data mining framework in this study manages relationships between banking organizations and their customers. Two typical data mining techniques - Neural Network and Association Rules - are applied to predict the behavior of customers and to increase the decision-making processes for recalling valued customers in banking industries. The experiments on the real world dataset are conducted and the different metrics are used to evaluate the performances of the two data mining models. The results indicate that the Neural Network model achieves better accuracy but takes longer time to train the model.

Keywords— Customer Relationship Management, Customer Knowledge Management, Data Mining, Neural Networks, Association rules

I. INTRODUCTION

Customer Relationship Management (CRM) is a modern management tool that, employs information technology such as database management, data analysis, and data mining to understand, target, and attract customers, with the objective of satisfying and retain them [1].

The Customer Knowledge Management (CKM) model has drawn attention recently through the convergence of both the technology-driven and data-oriented approach in CRM and the people-oriented approach in Knowledge Management (KM), with a view to exploiting their interaction potential [2,3,4] The expectation of this effort is to derive more comprehensive knowledge for customers; knowledge about customers, and knowledge from customers.

Data mining is defined as a process that uses mathematical, statistical, artificial intelligence and machine learning techniques to extract and identify useful information and subsequently gain knowledge from databases. Data mining algorithms have been widely used in range of research fields such as healthcare and medicine [5-7], sentiment analysis [8, 9], education [10] etc. The purpose of applying data mining in bank industry is to use the available data to retain its best customers and to identify opportunities sell them additional services.

Information technology tools, and explosion in banks’ customer data has improved and changed the managing relationships between banks and their customers. Data mining can be used in banks and finance organizations for decision making and forecasting. One of the most common learning models in data mining that predicts the future customer behaviours is classification. The prediction is done by the classification of database records into several predefined classes based on certain criteria. Neural networks, decision trees, naive Bayes, logistic regression, association rule, and SVM are the common tools used for classification [11].

In this research, a new Customer Knowledge Management (CKM) framework based on data mining is proposed. The neural networks and association rule mining are used as two classification techniques in this study. Furthermore, for the CRM applications in the banking domain, customer segmentation, prospecting and acquisition, security, profitability, risk analyses, strengths and weaknesses have been taken into consideration to show the performance of classification models. The Saman Bank customers dataset are used in the experiments. K-fold cross validation technique used to evaluate the predictive methods. The comparison of the performances of multilayer perception neural network and association rule classifier are provided in terms of effectiveness and efficiency.

The rest of the article are organized as follows. Section 2 is about related works from the literature. Section 3 describes the proposed data mining based bank CKM framework. section 4 covers experiments and results analysis, and finally section 5 concludes the article.

II. RELATED WORKS

A. Customer Relationship Management

The term CRM emerged in the information technology (IT) vendor community and practitioner community in the mid-1990s [12]. It is often used to describe technology-based customer solutions, such as sales force automation (SFA). In the academic community, the terms “relationship marketing” and CRM are often used interchangeably [13]. However, CRM is more commonly used in the context of technology solutions and has been described as “information-enabled relationship marketing” [14]. Zablal et al. [15] suggest that
CRM is “a philosophically-related offspring to relationship marketing which is for the most part neglected in the literature,” and they conclude that “further exploration of CRM and its related phenomena is not only warranted but also desperately needed.” [16].

CRM is a strategic approach that is concerned with creating improved shareholder value through the development of appropriate relationships with key customers and customer segments. CRM unites the potential of relationship marketing strategies and IT to create profitable, long-term relationships with customers and other key stakeholders. CRM provides enhanced opportunities to use data and information to both understand customers and co-create value with them. This requires a cross-functional integration of processes, people, operations, and marketing capabilities that is enabled through information, technology, and applications [12,16,17].

In Kano’s model, one-dimensional quality elements refer to conventional ideas about quality: the customer satisfaction is proportional to the functionality of the product, where less function leads to less satisfaction and vice versa. However, the quality elements that generate only customer satisfaction and no dissatisfaction are categorized as ‘attractive quality elements’, corresponding to the ‘motivator’ or ‘satisfier’ in Herzberg’s Motivation-Hygiene theory; while the quality elements that generate only customer dissatisfaction and no satisfaction are categorized as ‘must-be quality elements’, corresponding to the ‘hygiene factor’ or ‘dis-satisfier’ in Herzberg’s Motivation-Hygiene theory. The ‘must-be’ quality elements may coexist with ‘attractive quality elements’ in a certain product [18,19].

A Kano-CKM framework was proposed to address the improvement of issues in the management of customer knowledge as shown in Fig. 2. In the Kano-CKM model, the CKM process comprises four stages bolstered by Kano’s Method, and the implementation scheme is described as follows [19].

**B. Data Mining for CKM**

Data Mining (a.k.a. Knowledge Discovery from data) is a technique that examines information stored in large database or data warehouse and find the patterns or knowledge in the data that are not yet known and discovered patterns and knowledge can be used to predict the meaningful trends and relationships. In this regard, customer knowledge obtained via a CRM system is a valuable intellectual asset for a company to develop or improve products and services in order to meet or even exceed customers’ expectations. CRM
systems that collect information for customer knowledge are classified into three main categories [1]:

1. Operational CRM systems enhances the efficiency of a CRM process through service center management and marketing-automation like database marketing.
2. Analytical CRM systems evaluates knowledge of an individual customer’s attitude, needs, and values for cluster analysis. Data mining is a typical technique in this category.
3. Collaborative CRM systems synchronizes customer communication time through channels such as e-mail, the Internet, and/or the telephone.

In the literature most studies on KM and CRM are treated in separate research domains. However, lately their mutual synergy potential has drawn the attention of researchers in the field. By employing KM in an effort to help CRM to transcend from its original technology-driven and data oriented approach into a more people-oriented ‘customer knowledge management’ model or CKM model, it has already invoked a convergence of the two [3, 4] The CKM model emphasizes a bi-directional communication channel. This interaction with customers and customer knowledge management, set up strategies for how a company can develop attractive innovative products, or improve its services to win the satisfaction of its customers.

Fig. 3. Proposed CKM framework based on data mining

III. PROPOSED DATA MINING BASED BANKING CKM FRAMEWORK

A. Proposed Framework

The proposed data mining based CKM framework is shown in Fig.3. Understanding the product benefits, customers identifications, and their satisfactoriness forms the initial phase of any problem in data mining. A close study and management of customer identification and their knowledge will help to identify attract and retain effective customers in the domain. The next phase of data preparation helps in preparing the data by the processes of cleaning, attribute selection, data transformation for further building up of models and their evaluation. Model construction in the CKM framework is a major step in which effective model that satisfy the business requirements are constructed. These models help in predicting the behaviour of the customers.

Model evaluation measures the effectiveness of the model for enhancing their performance.

B. Neural Network and Association Rule Mining

Neural networks (NNs) are multi-layer networks of neurons that are used to classify things or make predictions. Among the input, output and hidden layers of neurons the actual computations of the network are performed in the hidden layer, where each neuron sums its input attributes \( x_i \) after multiplying them by the strengths of the respective connection weights \( w_{ij} \). The activation function (AF) of this sum gives the output \( y_j \) and sigmoid function is the AF used in the experiment [20].

\[
y_j = f(\sum w_{ij} x_i)
\]

Back-propagation (BP) learning is the most common training technique used for NNs. The sum of squared differences between the desired and asset value of the output neuron's E is defined as:

\[
E = \frac{1}{2} \sum (y_{dj} - y_j)^2
\]
Where \( y_j \) is the output of a neuron \( j \) whose desired value is \( y_d \). Weights \( w_{ij} \) in equation (1), are adjusted to finding the minimum error \( E \) of equation (2) as early as possible. The difference between the network outputs and the desired ones is reduced by the application of weight correction by BP. The neural networks help in learning and reducing the future errors. Good learning ability, fast real-time operation, less memory demand, analysis of complex patterns is some of the advantages of NNs and the disadvantages include high-quality data requirement of the network, and careful selection of parameters.

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. Association Rule Association rule mining, one of the most important and well researched techniques of data mining, was first introduced in [21,22]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories.

Let \( I = I_1, I_2, \cdots, I_m \) be a set of \( m \) distinct attributes, \( T \) be a transaction that contains a set of items such that \( T \subseteq I, D \) be a database with different transaction records \( T_x \). An association rule is an implication in the form of \( X \Rightarrow Y \), where \( X, Y \subseteq I \) are sets of items called item sets, and \( X \cap Y = \emptyset \). \( X \) is called antecedent while \( Y \) is called consequent, the rule means \( X \) implies \( Y \). There are two important basic measures for association rules, support (s) and confidence (c). The two basic parameters of Association Rule Mining (ARM) are: support and confidence. Support (s) of an association rule is defined as the percentage/fraction of records that contain \( X \cup Y \) to the total number of records in the database. The count for each item is increased by one every time the item is encountered in different transaction \( T \) in database \( D \) during the scanning process. It means the support count does not take the quantity of the item into account.

Support (s) is calculated by the following:

\[
Support(XY) = \frac{\text{Support count of } XY}{\text{Total number of transaction in } D}
\]  

Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain \( XY \) to the total number of records that contain \( X \), where if the percentage exceeds the threshold of confidence an interesting association rule \( X \Rightarrow Y \) can be generated.

\[
Confidence(X|Y) = \frac{Support(XY)}{Support(X)}
\]

### IV. EXPERIMENTS AND RESULTS

#### A. Study Dataset

The dataset used for experiments in this paper, contains results of the Saman bank customers data. Saman Bank is one of the Iranian private banks which established recently [23].

Data includes answers of 124 Saman Bank customers and were collected by questionnaires. From the overall 255 filled questionnaires only 124 questionnaires were completely answered all questions and then were selected to use in this study. The questionnaire includes 21 questions which cover the bank services and customers satisfaction rates. The variables of the questionnaire are: age, average yearly balance in Iranian Rials, job type, marital status, gender, education, type of account, type of loan, satisfactory of electronic bank services, satisfactory of security, satisfactory of bank facilities, satisfactory of complaint reports, satisfactory of banking service, satisfactory of bank informativity. All these variables are used as features for prediction of customer behaviour in two classes (satisfactory/non-satisfactory). The details of the features are described in Table I.

<table>
<thead>
<tr>
<th>Feature Number</th>
<th>Feature Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>age</td>
<td>Numeric</td>
</tr>
<tr>
<td>2</td>
<td>average yearly balance</td>
<td>Numeric</td>
</tr>
<tr>
<td>3</td>
<td>job type</td>
<td>Numeric</td>
</tr>
<tr>
<td>4</td>
<td>marital status</td>
<td>Numeric</td>
</tr>
<tr>
<td>5</td>
<td>education</td>
<td>Numeric</td>
</tr>
<tr>
<td>6</td>
<td>type of account</td>
<td>Numeric</td>
</tr>
<tr>
<td>7</td>
<td>gender</td>
<td>Categorical</td>
</tr>
<tr>
<td>8</td>
<td>type of loan</td>
<td>Categorical</td>
</tr>
<tr>
<td>9</td>
<td>satisfactory of electronic bank services</td>
<td>Categorical</td>
</tr>
<tr>
<td>10</td>
<td>satisfactory of security</td>
<td>Categorical</td>
</tr>
<tr>
<td>11</td>
<td>satisfactory of bank facilities</td>
<td>Categorical</td>
</tr>
<tr>
<td>12</td>
<td>satisfactory of complaint reports</td>
<td>Categorical</td>
</tr>
<tr>
<td>13</td>
<td>satisfactory of banking service</td>
<td>Categorical</td>
</tr>
<tr>
<td>14</td>
<td>satisfactory of bank informativity</td>
<td>Categorical</td>
</tr>
</tbody>
</table>

### B. Experiment Setup

All experiments were performed using weka tool and were conducted in windows 10 with Intel Core i7 processor. The Waikato Environment for Knowledge Analysis (Weka) is a machine learning toolkit used extensively for research, education and projects. Weka is introduced by Waikato University, New Zealand and is open source software written in Java (GNU Public License). It consists of collection of machine learning algorithms and tools for data mining tasks such as data pre-processing or data preparation, classification, association rules, clustering, regression, forecasting and visualization and is well suited for developing new machine learning schema.

In this work, we build two distinct DM classifier models: Neural Networks (NN) and Association Rule (AR). For all the two models test mode of tenfold cross validation was used. NN uses back propagation to classify the instances. The nodes are all sigmoid except for when the class is numeric. We set
the number of hidden layers using the heuristic $a = \text{round} \left( \frac{M}{2} \right)$ where $M$ is the sum of attributes and classes. Other network parameters were set as follows: learning rate 0.2, momentum $0.1$, training time 100ms.

During the modelling phase we successfully tested the two models, NN and AR using the Weka tool. Based on the response, two classes were obtained, those which responded positive and those responded negative.

C. Evaluation Metrics

Evaluating of a machine learning algorithm is an essential part of any project. Most of the time classification accuracy is used to measure the performance of a model, however it is not enough to truly judge a model. Classification accuracy and validation accuracy are the ratio of number of correct predictions to the total number of input samples in both training and testing sets. Different evaluation metrics were measured in this research to measure the effectiveness of the proposed method and hypotheses such as ROC, accuracy, TPF, FPR, F-measure, precision, recall, sensitivity, and specificity.

A confusion table as shown in Table II has two rows and two columns that reports the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

Accuracy: the percentage of instances classified correctly into a given category in relation to the total number of instances tested:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

The F score is used to measure a test's accuracy, and it balances the use of precision and recall to do it. The F score can provide a more realistic measure of a test's performance by using both precision and recall. F-measure and precision are calculated based on FP, TN, FN, and TP.

ROC area (area under the ROC curve): In a Receiver Operating Characteristic (ROC) curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test.

D. Results and Discussion

The results of our experiment for automatically classifying a given dataset are summarized in Table III. Which shows values for two different classifiers. For each method the classification accuracy (amount of correctly classified instances), true positive rate (TPR) (the proportion of actual positives which are correctly identified as such), false positive

![Fig. 4. The comparison between different evaluation metrics for NN and AR.](image-url)

TABLE II. CONFUSION MATRIX

<table>
<thead>
<tr>
<th>Actual: No</th>
<th>Predicted: No</th>
<th>Predicted: Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Negatives (TN)</td>
<td>False Positives (FP)</td>
<td></td>
</tr>
<tr>
<td>False Negatives (FN)</td>
<td>True Positives (TP)</td>
<td></td>
</tr>
</tbody>
</table>

TABLE III. RESULTS OF COMPARISON OF CLASSIFIERS, AVERAGE OVER 10 RUNS.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy (%)</th>
<th>TPF</th>
<th>FPR</th>
<th>ROC</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>97%</td>
<td>0.87</td>
<td>0.042</td>
<td>0.92</td>
<td>1500</td>
</tr>
<tr>
<td>AR</td>
<td>94.3%</td>
<td>0.84</td>
<td>0.052</td>
<td>0.90</td>
<td>12</td>
</tr>
</tbody>
</table>

TABLE IV. F-MEASURE, PRECISION, RECALL, SENSITIVITY, SPECIFICITY METRICS RESULTS.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>F-Measure</th>
<th>Precision</th>
<th>Recall</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>88%</td>
<td>0.83%</td>
<td>85.5%</td>
<td>84%</td>
<td>87%</td>
</tr>
<tr>
<td>AR</td>
<td>82%</td>
<td>78.5%</td>
<td>81%</td>
<td>79%</td>
<td>83.4%</td>
</tr>
</tbody>
</table>

TABLE V CLASSIFICATION OF 124 INSTANCES IN THE DATASET

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Correctly Classified Instants</th>
<th>Incorrectly Classified Instants</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>120</td>
<td>4</td>
</tr>
<tr>
<td>AR</td>
<td>116</td>
<td>8</td>
</tr>
</tbody>
</table>
V. CONCLUSION AND FUTURE WORK

In this paper, an efficient CKM-data mining framework for the prediction of customer behavior was proposed. Two representative data mining techniques - Neural network and association rules – were applied in the new CKM base data mining framework for bank customer behavior predicting in the real world. Our findings indicate that both models attain good predicting performance and data mining techniques can be used to improve customer behavior understanding and predicting. The best model that achieves high predictive performance was Neural Network with accuracy rate of 97%. However, Neural Network takes longer time to train the model.

In the future, this work can be extended to use other new models like Neuro fuzzy classifiers, Ensemble of classifiers and so on in order to improve the predicting capacity. Also, the other large banking datasets should be used for testing the proposed new framework.

REFERENCES


