A Data Mining Approach for Saving Service Selection

Richa Sapra  
Lovely Professional University  
Phagwara  
sapra.richa20@gmail.com

Sheveta  
Lovely Professional University  
Phagwara  
shevetavashisht@ymail.com

Mr. Mandeep Singh  
Lovely Professional University  
Phagwara  
mandeep.13742@lpu.co.in

Abstract

Savings were available for use whenever needed, but it also involved the risk of loss by theft, robbery and other accidents. One can invest the money after consulting the post office employees or going through the features of the available saving services. But most of the clients face the problem of selecting the best saving service for investment of their money as number of saving services exist. Moreover, it is a time consuming process to select one of the best saving service by the use of traditional method. Various data mining techniques can be applied to develop such a method for selecting the best saving service. This paper presents an application on which data mining technique have been applied for clustering and classification.

1. Introduction

With the development of information technology and computer science, high-capacity data appear in our lives. In order to help people analyzing and digging out useful information, the generation and application of data mining technology seem so significance [1]. People were in need of a place where money could be saved safely and would be available, whenever required. Banks and post offices are such places where people can deposit their savings with the assurance that they will be able to withdraw money from the deposits whenever required. Various saving services provide the basis of protection and benefits on the money invested under the particular scheme [2]. For this, one can either consult various post offices or bank employees or use the Internet sites to gather the large amount of knowledge and get the guidance for investment. While the use of bank employees and Internet to have the complete knowledge about the best saving service is time consuming and tedious process. The manual extraction of patterns from data has occurred for centuries. The proliferation, ubiquity and increasing power of computer technology has increased data collection, storage and manipulations. As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries such as neural networks, clustering, genetic algorithms, decision trees and support vector machines. Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns. A primary reason for using data mining is to assist in the analysis of collections of observations of behavior. So it can be defined as “Data Mining is the process of discovering new patterns from large data sets involving methods from statistics and artificial intelligence but also with database management”.

A lot of work has been done on banking credit system [3] [4]. But none of the researchers have developed system for available saving services of banks and post offices. These days an abundant information data can be accumulated from banks and post offices about the available saving services. Furthermore, banking itself have some peculiarities such as complexity and variety of information which make the information dataset tend to large-scale, dynamic and uncertain. This huge amount of dataset is difficult to manage and it is very difficult to extract information and take decision from the huge amount of dataset.

Over the years, a variety of algorithms for classification and taking decision in large databases have been developed. Data mining algorithms have been used extensively in different applications. These algorithms are used to analyze business, commerce, science, engineering and security. Data mining can
describe and predict different datasets or knowledge by different technologies. So, we can deal with these banking peculiarities effectively and selecting the best saving service from the available services will naturally be a new research point. The main objective of this paper is to deploy a data mining model by which the available saving services can be segmented and classified. The rest of this paper is organized as follows. Section 2 describes the basic idea related to the work. Section 3 describes the proposed system architecture and elaborates the algorithm which includes three important parts: Clustering, Classification and prediction. Then the flow out of the algorithm has been given out. The section 4 describes the implementation and section 5 includes the conclusion and future work.

2. Basic Idea

Clustering and decision tree are two of the mostly used methods of data mining which provide us much more convenience in researching information data. This paper presents the proposed method in which both clustering and Decision tree algorithms are used. Clustering is the old and widely used method of Data mining. In clustering, the records which are alike are grouped or clustered together and put into the same grouping. Clustering is the unsupervised learning technique in the sense that when they are run, there is no particular reason for the creation of the models the way that the classification and prediction can be done [5]. Classification is also an important task in data mining. Its purpose is to set up a classifier model and map all the samples to a certain class which can provide much convenience for people to analyze data further more. Classification belongs to directed learning, and the main methods include decision tree, Bayesian classification, neural network, genetic algorithm and rough set etc.

2.1. K-Means Clustering Algorithm

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation’s proximity to the mean of the cluster. The cluster’s mean is then recomputed and the process begins again. Here’s how the algorithm works:

1. The algorithm arbitrarily selects k points as the initial cluster centers (“means”).
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters [6].

2.2. C4.5 Decision Tree Algorithm

Decision tree is one of the important analysis methods in classification. It builds its optimal tree model by selecting important association features. While selection of test attribute and partition of sample sets are two crucial parts in building trees. Different decision tree methods will adopt different technologies to settle these problems. Traditional algorithms include ID3, C4.5, CART, CHAID etc. C4.5 Decision Tree Algorithm: C4.5 is the enhancement of ID3 algorithm that improves the performance. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. C4.5 builds decision trees from a set of training data in the same way as ID3 using the concept of information entropy. The training data is a set \( S = \{ s_1, s_2, ..., s_i \} \) of already classified samples. Each sample \( s_i \) is a vector \( x = (x_1, x_2, ..., x_n) \) where \( x_1, x_2, ..., x_n \) represent attributes or features of the sample. The training data is augmented with a vector \( C = (c_1, c_2, ..., c_n) \) where \( c_1, c_2, ..., c_n \) represent the class to which each sample belongs [7]. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurses on the smaller sub lists. This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a
decision node higher up the tree using the expected value of the class.

- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value [7]

3. Proposed System Architecture

These days, whenever data mining algorithms are discussed, people usually talking about either decision trees or neural networks. Neural Network algorithms have been of greater interest through data mining technology [8]. The basic idea of the proposed work is to select the best saving service from the dataset. For this two phase process which combines K-means for clustering and Decision Tree (C4.5 algorithm). These two algorithms are data mining algorithms for clustering and classification respectively so that predictions can be made. On basis of the results user may select the saving service. The specific approach is as follows:

- First of all the dataset have been created by sampling the data and make it standardized.
- The standardized dataset is used as Input sample.
- The K-means clustering algorithm is applied and the clusters are formed. The groups are formed after generation of the matrix. By clustering method all the saving services having same features are grouped together.
- Apply the Decision Tree algorithm C4.5 to the cluster which input sample belonging to.
- On the basis of the classified data, decisions have been made.

![Figure 1. Specific Flow of Algorithm](image)

4. Practical Experiment

4.1. Experimental Environment

The main software package used in our experiments is Weka (Waikato Environment for Knowledge Analysis). It is a free, non-commercial, powerful open source Java-based machine learning, and data mining software, developed at the University of Waikato in New Zealand. Weka and its source code can be publicly available online at Http://www.cs.waikato.ac.nz/ml/weka.

Implementation on classification and selection of saving service has following phases. First of all the attributes have been collected with the help of Internet and after visiting the post offices and banks. The dataset for various schemes has been generated including the important attributes such as: Interest on
4.2. Experiment and Result

**Table 1: Dataset for the model**

From the sample data as shown in table 4.2 the services S1, S2, S3 and S4 belong to cluster 1, cluster2, cluster3 and cluster4.

Now when the complete data is clustered into four groups and there is a need to classify the services, classification needs to be done. The classification algorithm J48 (C4.5) is used to training the model. The output of the clusters is fed to WEKA tool for classification.

Following figure shows the results in ARFF format which is generated after the classification.

**Figure 3. The classifier Output**

After uploading the data file, the clustering is performed. In this the K-means clustering is used. The Euclidean distance is used as a distance function to form the centroid of the clusters. The maximum iteration number is 500 and number of cluster to be formed is set at 4, the exact formation of the clusters and the centorid of the two clusters. Cluster 0 contains the 16%, cluster 1 contains the 52%, cluster2 contains 13% and cluster3 contains 19% of the total dataset.
5. Conclusion and Future Work

In this paper combination of two data mining algorithms proposed. This method is very user friendly and less time consuming. This research uses the combination of two very popular Data mining techniques, KMeans for clustering and C4.5 (J48 in Weka) is used to classify the Saving Services. Through this model, the user who wants to invest his money may easily take the decision as services are classified to Good, bad, Medium and Very Good level.

The accuracy of the proposed work is 95.04 % and it efficiently classifies the services. This research may extend for those users who have their preferences to select the particular features of the scheme in which they can take decision to do investment according to their needs.

References


