ISSN No. 0976-5697

Volume 4, No. 9, July-August 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Approach for Labeling the Class of Credit card Customers via Clustering Method in Data Mining

Mizanur Rahman
M. Sc Student, Computer science and engineering
Islamic University
Kushtia, Bangladesh
milon06cse@yahoo.com

Sujit Kumar Mondal
Associate Professor, Computer science and engineering
Islamic University
Kushtia, Bangladesh
sujit_iu@yahoo.com

Abstract: Data mining or knowledge discovery is the process of discovering hidden patterns in large data sets. Clustering is one of the techniques of data mining. Credit cards are growing as a popular medium of transaction which is flexible, secure and much safer from theft than travel with cash and also a promising area for buying and sales. In this paper, most well-kwon clustering technique k-means is used to classify the real life data of credit card customer in Bangladesh. The goal of this research is to avoid default customer and find criteria of profitable and long lasting customer for credit card. To understand criteria of profitable customer, output of clustering is used to find valuable pattern.

Keywords: Data mining; clustering; k-means; credit card

I. INTRODUCTION

Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions [1]. Businesses collect large amounts of information on current and potential customers [2]. Financial institutions like banks are interest to catch the competitive market of card, for providing new product or much facility to their customer and to attach new customer to banking with them, for this reason banks need to analyze data of customer. In this paper, customer data is analyzed, clustered and finally retrieve useful information from it.

II. CREDIT CARD

Credit card is another name of short term loan or advance, so it is important to find safe and profitable credit card customer. Credit scoring has become very important issue due to the recent growth of the credit industry, so the credit department of the bank faces the huge numbers of consumer's credit data to process, but it is impossible analyzing this huge amount of data both in economic and manpower terms [3]. Many of these proposed models can only classify customers into two classes “good” or “bad” ones. Several single and hybrid data mining methods are applied for credit scoring problem [4], [5].

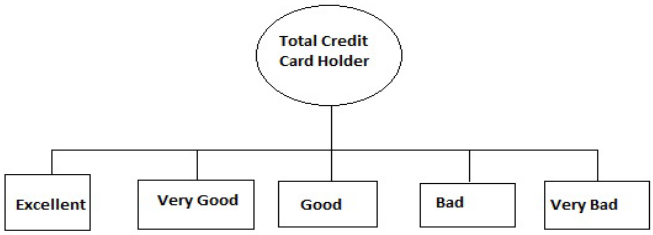

Fig1: Clusters of credit card holder

In this research, we classify the total card holder into five classes as: Excellent, Very good, Good, Bad, Very bad according to their age, occupation type, position in organization, salary, billing history and default record.

III. DATA MINING

Data mining methods are algorithms that are used for building models and for finding patterns in data. Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses [1]. Data mining is primarily used today by companies with a strong customer focus - retail, financial, communication and marketing organizations. The following list shows the most common data mining tasks [1].

- Description
- Estimation
- Prediction
- Classification
- Clustering
- Association

Data mining is having lot of importance because of its huge applicability. It is being used increasingly in business applications like market research, consumer behavior, direct marketing, bioinformatics, genetics, text analysis, e-commerce, customer relationship management and financial services for understanding and then predicting valuable data, like consumer buying actions and buying tendency, profiles of customers, industry analysis, etc. In this thesis, clustering task is used to analyze customer data.

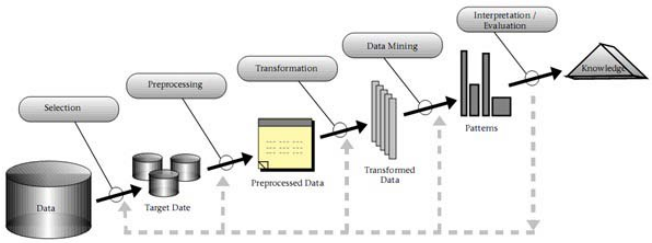

Fig2: Data mining process

IV. CLUSTERING

Clustering is an important unsupervised classification technique used in identifying some inherent structure present in a set of objects [6],[7]. Clustering refers to the grouping of records, observations, or cases into classes of similar objects. A cluster is a collection of records that are similar to one another and dissimilar to records in other clusters [1]. Clustering refers to the process of grouping samples. Formal, mathematical definition of clustering, as started following: let X□ R^{mxn} a set of data items representing a set of m points X_i in R^n. The goal is to partition X into K groups C_k such every data that belong to the same group are more alike than data in different groups. Each of the K groups is called a cluster [8]. Clustering is one kind of data mining task in which data of database are grouped into several classes [9]. Clustering can be used to segment customers into a small number of groups for additional analysis and marketing activities [10]. Clustering can be classified into two major types, Hierarchical and Partitioning clustering.

A. *Hierarchical clustering*

Hierarchical clustering algorithms recursively find nested clusters either in agglomerative mode (starting with each data point in its own cluster and merging the most similar pair of clusters successively to form a cluster hierarchy)or in divisive (top-down) mode(starting with all the data points in one cluster and recursively dividing each cluster into smaller clusters)[11].

Hierarchical Clustering algorithms:
The Single-Linkage Algorithm
The Average-Linkage Algorithm
The Complete-Linkage Algorithm
Ward's Method

B. *Partitioning clustering*

Partitioning clustering algorithms find all the clusters simultaneously as a partition of the data and do not impose a hierarchical structure [11].

Partitioning Clustering algorithms:
Forgy's Algorithm
K-means algorithm
Isodata Algorithm

In this research work, k-means clustering algorithm is used to grouping the entire customer into five groups.

V. K-MEANS CLUSTERING

Let $X=\{x_i\}$, i= 1,...,n be the set of n dimensional points to be clustered into a set of K clusters, $C=\{C_k, k=1,....,k\}$. K-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized [3]. Let, μ_k be the mean of cluster C_k. the squared error between μ_k and the points in cluster C_k is defined as

$$J(c_k) = \sum_{x_i \in c_k} ||x_i - \mu_k||^2 \quad (1)$$

The goal of k-means is to minimize the sum of the squared error over all K clusters,

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} ||x_i - \mu_k||^2 \quad (2)$$

Besides the data, input to the algorithm consists of k, the number of clusters to be constructed. The k-means algorithm differs from other partitioning algorithm in that the centroids of the clusters are recomputed as soon as a sample joins a cluster. Also, unlike Forgy's algorithm which is iterative, the k-means algorithm makes only two passes through the data set.

VI. METHODOLOGY

Our goal is to classify 19000 customer data with a pre specified centroid for each cluster. After creating the final clusters we interpret what are the criteria of a customer having bad or good credit risk. To do this clustering process we divided our work as following steps:

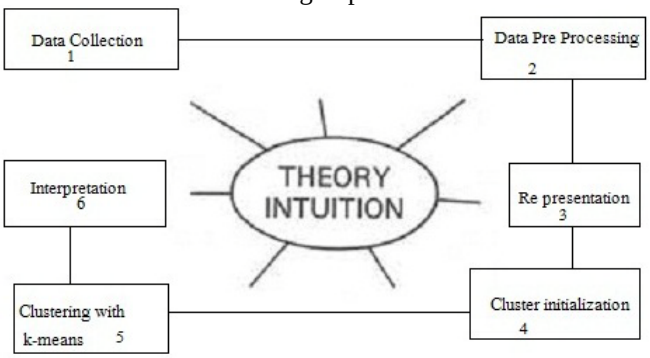

Fig3: Clustering Methodology

A. *Data collection*

Data collection is the first part of clustering work. We collect data of VISA credit card from Dutch-Bangla Bank Limited, Bangladesh. our data set contain records of 19000 current VISA credit card customer with Id, age, occupation, Salary/income, types of organization he/she work, position in company, credit limit, bill of last six month and number of month he/she unable to pay bill.

B. *Data preprocessing*

Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Preprocessing is the primary step of clustering because the raw data contained in databases is unprocessed, incomplete, and noisy. To be useful for data mining purposes, the databases need to undergo preprocessing, in the form of data cleaning and data transformation [1].

Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this dirty data. The First task of data preprocessing is the data cleaning, there are many inconsistent, old and garbage value of fields exist in data base. In Credit data base, Many of the customer has more than one credit limit because this limit is revised several time for the increment of customer's salary or monthly income or credit limit may be decreased for his/her irregular payment.

Missing data or missing values occur when no data value is stored for the variable in an observation. Missing data is a problem that continues to plague data analysis

methods [1]. Many customers does not provide their information like age, marital status etc or many times customer relationship officer missed the field. A common method of handling missing values is simply to omit from the Analysis the records or fields with missing values [1]. But sometimes this is very harmful because the pattern may be lost. In this paper, missing value is replaced by field mean.

Data transformation allows the mapping of the data from its given format into the format expected by the appropriate application. This includes value conversions or translation functions, as well as normalizing numeric values to conform to minimum and maximum values. In banking database there is large range of some field like age, income, loan amount. In this paper, Min-Max Normalization is used to transform data in a range from 0 to 1.

Min-max normalization works by seeing how much greater the field value is than the minimum value min(X) and scaling this difference by the range [1]. That is,

$$X* = \frac{\mathrm{X} - Min(X)}{Range(X)} \quad (3)$$

C. *Data Analysis and Re presentation*

The data set contains 15 variables of whether worth of information about 19000 current visa credit card customers. The variables are as follows:

- ID: Integer value, Unique Identity number of individual customer, actually it is customer id in database
- Age: integer value, how old the customer is
- Sex: Integer value, 1 represent Male and Female 2
- Monthly income/salary: Double value, Customer's monthly income or gross salary.
- Occupation: Text type, it may be Job, Business etc
- Position: Text type, Designation in Occupation, like Proprietor, Chairman, Professor etc.
- Company: text type, Type of organization where he/she work, it may be Government or public, NGO, small business or large business etc.
- Marital status: Married(2) or not(1)
- Six month bill: Six double value variable to represent last six month billing statement of customer.
- Defaulter History: Integer value, how many month the customer fail to pay bill.

D. *Cluster initialization*

Assuming seeds for clusters by divide the feature ranges in 5 groups for five clusters according to credit risk calculation matrix of bank as
Age:
Excellent: 50-62
Very good: 40-50
Good: 30-40
Bad: 63-65& 26-30
Very Bad: More than 65 and less than 25
Income range:
Excellent: 80 or above
Very good: 60-80
Good: 40-60
Bad: 25-40
Very bad: less than 25
Bill range:
Excellent: 71-100% of limit
Very good: 51-70% of limit
Good: 41-50% of limit
Bad: 21-40% of limit
Very bad: 0-20% of limit
Default history range:
Excellent: 0
Very good: 1
Good: 2
Bad: 3 month
Very bad: 4 or above month
Company nature Range:
Excellent: Multinational Company, limited, joint venture, Top ranked bank
Very good: university, Bank, defense,
Good: International NGO, Govment first class, College
Bad: Small agency, School, coaching
Very Bad: Local NGO, Small Company, Small business
Position on company range:
Excellent: MD, EVP-DMD, Chairman, Minister and so on
Very Good: AVP or above, Professor, Dr, AGM, SP, DC
Good: SO, SEO, in charge, Asst. professor, grade-11 government employee, Proprietor of large business
Bad: Grade-9, Lecturer, officer, Proprietor of medium business
Very Bad: 3rd-4th class Worker, student, retired person, Serious patient, Proprietor of small business, Unemployed
Seed format for cluster
Fv [age, sex, income, position, company type, marital status, bill1, bill 2, bill 3, bill 4, bill 5, bill 6, default]

E. *Clustering with k-means*

Data set is clustered by k-means algorithm, equation () is used to find minimum squared error. There are two step in K-means algorithm [12] as

1. Begin with k clusters, each consisting of one of the first k samples. For each of the remaining n-k samples, find the centroid nearest it using equation 1 and 2. Put the sample in the cluster identified with this nearest centroid. After each sample is assigned, re-compute the centroid of the altered cluster.
2. Go through the data a second time. For each sample, find the centroid nearest it. Put the sample in the cluster identified with this nearest centroid. (During this step, do not re-compute any centroid.)

F. *Interpretation*

Customers having age about 42 with about 50 thousand monthly income, used about 65% of credit limit and married with no default history are excellent credit risk. Customers having age about 50 with about 60thousand monthly income, used about 40% of credit limit and married with only no default history are very good credit risk. Customers having age about 20 with about 15 thousand monthly income, used about 10% of credit limit and unmarried with 4 pending bill are very bad credit risk.

VII. RESULT AND DISCUSSION

Compared to traditional credit risk matrix calculation, k-mens clustering shows customer having much more

salary or monthly income are not excellent all the time. K-mans clustering result provide more focus on the use of credit limit because if customer does not use his/her credit limit the money became idle which create low benefits in banking business. K-means clustering show that employeed females are very good credit risk than male. Besides some client record resides on the one cluster are similier to other but they are in different cluster because this research use credit limit used by customer. If salary of one customer is same to other but they have different credit risk due to default history and record of card used. Customer density is very low in excellent and very good class. Future research will be on the transaction history of customer to detect card fraud on ecommerce buying and sales.

Table A: A small portion of credit card customer data

ID	Company	Age	Position	Marital Status	Credit Limit	Salary	Month1	Month2	Month3	Default
348	DUTCH-BANGLA BANK	27	OFFICER	1	60000	30000	30566.67	1156.46	53383.54	0
433	NIPRO JMI COMPANY	35	HEAD OF PMD	2	75000	37500	1437.5	937.5	15201	0
336	EXPRESS SYSTEM	37	MANAGER	2	100000	33000	862.5	362.5	37.5	4
9837	DUTCH-BANGLA BANK	27	OFFICER	1	60000	30000	35850.42	20200	16045.83	1
3995	B M STORE	40	PROPRIETOR	2	200000	66000	181446.5	181145	181225	0
952	BD FINANACE	36	FAVP	2	80000	40000	862.5	30297.5	31257.5	0
763	THE CITY BANK	37	VP	2	115000	32000	1437.5	4917.5	2.5	3
964	TARASIMA APPARELS	31	MANAGER	2	100000	43000	862.5	362.5	18369.14	0
5465	SOUTH EAST TEXTILE	30	OFFICER	2	30000	15000	0	9340	0	0
6990	ANWAR LANDMARK	34	MANAGER	2	80000	40000	58361.13	61307.5	60235.22	0
6037	APOLLO HOSPITAL	32	SMO	2	80000	40000	862.5	362.5	45.4	2
342	DESH GENERAL INSURANCE	51	D M D	2	225000	75000	4007.9	0	1449	0

Table B: Updated cluster centroid with some customer id in corresponding cluster

Cluster	*Cluster centroid*	*id of Customer*
Excellent	{42,1,52,3,4,2,65,0}	{5,58,69,121,205,308,584,2108,3203,8360,9612,10205,13053,15052,18001,…..}
Very Good	{50,2,43,4,3,2,50,0}	{41,687,925,1042,1141,1584,10625,10764,10585,11269,12202,13569,15641,.....}
Good	{35,1, 30,3,2,2,35,1}	{14,125,451,565,683,2251,2143,2356,2471,2920,3003,3243,3567,4891,9924,.....}
Bad	{25,1,25,4,3,1,25,2}	{86,95,601,548,659,845,932,2365,5896,6587,7124,8218,120810,15282,17216,.....}
Very Bad	{16,1,15 ,4,3,1,10,3}	{2,83,211,538,684,914,2635,9087,107124,120358,12581,13213,18210,18321,.....}

Fig 4: Output with centroid mark

VIII. ACKNOWLEDGMENT

We would like to thank the officers of Cards Operation Division, Dutch-bangla bank limited for providing data and supporting my research in data clustering and classification. I am grate full to A.H.M. Sajedul Hoque, Assistant Professor, Dept. of Computer Science & Engineering, Hamdard University, Bangladesh for his valuable help.

IX. REFERENCES

[1] Daniel T. Larose, "Discovering knowledge in data : An Introduction to data mining ", John Wiley & Sons, Inc, 2005.

[2] M.R.anderberg, “Cluster analysis for application”, academic Press, New York 1973.

[3] Neelamadhab Padhy , Dr. Pragnyaban Mishra ,and Rasmita Panigrahi, "The survey of data mining applications and feature scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[4] Zurada, J., and Lonial, S., "Comparison of the Performance of Several Data Mining Methods for Bad Debt Recovery in the Healthcare Industry." the Journal of Applied Business Research, 21(2), 37-53. 2005.

[5] Kirkos, E., Spathis, C., and Manolopoulos., Y., 2007, "Data Mining techniques for the detection of fraudulent financial statements." Expert Systems with Applications 32(4), 995-1003.

[6] M.R.anderberg, “Cluster analysis for application”, academic Press, New York 1973.

[7] A.K. Jain, R.C. Dubes, "Algorithms for clustering data", Prentice-Hall, Englewood Cliffs, NJ,1988.

[8] T. Graepel, "Statistical physics of clustering algortihms", Technical Report 171822, FB Physik, Institut fur Theoretische Physic, 1998.

[9] Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth "From Data Mining to Knowledge Discovery in Databases", AI MAgazine, 1996.

[10] M.R.anderberg, “Cluster analysis for application”, academic Press, New York 1973.

[11] Anil K. Jain,” Data clustering: 50 years beyond k-means”, ELSEVIER, Pattern Recognition Lett.(2009).

[12] Igor Aleksander and Helen Morton, "An Introduction to Neural Computing", Published by Chapman and Hall, London,1991.