Modified Approach for Classifying Multi-Dimensional Data-Cube Through Association Rule Mining for Granting Loans in Bank

Dr. K.Kavitha
Assistant Professor, Department of Computer Science
Mother Teresa Women's University, Kodaikanal

Abstract *In this paper, modified Approach for classifying Multi-dimensional data cube is constructed. It explores data cubes in large Multi-Dimensional Schema. Numerical and Nominal attributes are categorized based on Principal Component Analysis. Semantic relationships are identified by applying Multi-dimensional scaling. Additionally, AR is integrated for finding the inserting measures. Many algorithms have been proposed for applying Multi-dimensional schema. But still some difficulties to category wise the integrated rules. The proposed approach suggested a new idea for categorizing the rules by using bank loan detect. This method provides accurate prediction and consumes less time than existing method.*

Keywords---*Association Rules, Datacubes, Data Mining, Multidimensional Schema, Information Gain*

I. INTRODUCTION

Data mining facilitates the discovery of unrevealed trends from large voluminous data sets. Data warehousing provides an interactive analysis of data through the use of different data aggregation methods. Data warehousing contributed key technology for complex data analysis, automatic extraction of knowledge from wide data repositories and decision support. Recently, there has been an increased research going on to integrate those two technologies. Moreover this work is concentrated on applying the data mining technique as a front end technology to a data warehouse to extract trends and rules from the data repository in data warehouses.

A huge range of data mining techniques has made significant improvements to the field of knowledge discovery in various domains. For example in the banking sector, these methodologies are used for loan payment prevision, classification of customers for targeted marketing, customer credit policy analysis, detection of money laundering schemes and other financial crimes. The banking database should control credit management thoroughly. Loan sanctioning requires the usage of huge data and significant processing time. To sanction the loan to customers, the bank needs to take some kind of precautions such as performance of the customer firm by analyzing the previous year's financial statements. The major tool used in multidimensional analysis in a data warehousing is the use of data aggregation and exploratory techniques that forms a part of (OLAP). The traditional OLAP technique is limited to detect hidden association between the items side in a data warehouse. So a lot of research undergoes to extend the OLAP technique to anticipate the future events. In this paper the capability of OLAP is extended to detect the hidden association and forecast the future events based on the historical data driven from the multidimensional schema.

In this paper, a novel approach called datacubes association rule algorithm is proposed across multidimensional datacubes. This method is based on information gain to detect and rank the most informative dimensions among the nominal variables. The application of Principal Component Analysis extracts the informative datacubes at different level of data abstraction. With the help of objective measures interesting rules are discovered.

The rest of the paper is organized as follows. Section II presents a description about the previous research which is relevant to the design and analysis of multidimensional schema. Section III describes about the existing model DCAR and its limitations. Section IV involves the detailed description about the proposed method. Section V presents the performance analysis. This paper concludes in Section VI.

II. RELATED WORK

This section deals with the works related to the association rules in data mining and multidimensional schema. *Usman, et al* proposed a methodology that selects a subset of informative dimension and fact variables from an initial candidate set. The experimental results of this method were

conducted on three real world datasets, which was extracted from the UCI machine learning repository. The knowledge discovered from the schema was more diverse and informative than the standard approach of the original data [1]. *Pears, et al* presented a generic methodology which incorporated semi-automated knowledge extraction methods to provide data-driven assistance towards knowledge discovery. A binary tree of hierarchical clusters were constructed and numeric variables were annotated. Three case studies were performed three real-world datasets from the UCI machine learning repository in order to validate the generality and applicability [2].
[1] Liu et al proposed a kind of personal financial recommendation system based on association rules. The data cubes were generated based on the customer's financial information, then the multi-dimensional association rules were generated. This system was more intelligent and personalized in solving the complexity in customizing financial service [3]. *Chiang* proposed an improved model to mine association rules of customer values. Ward's method was adopted to partition the online shopping market into three markets. The supervised Aprori algorithm was employed to create association rules [4].
Herawan and Deris presented an alternative approach for mining regular association rules and maximal association rules from transactional datasets. The transactional dataset was transformed into a Boolean-valued information system. The notions of regular and maximal association rules between two sets were defined. Also, support, confidence, maximal support, maximal confidences were defined using soft set theory [6]. *Kumar and A. Chadha* presented a case study of a university that hopes to improve the quality of education. So that association rule discovery was used to predict the accurate results [7]. *Romero, et al* explored the extraction of rare association rule from a Moodle system. Some relevant results were obtained and compared with the rare association rule mining algorithm [8]. *Zhu and Li* presented an association rule mining algorithm Ex- Apriori which was based on the predicate path graph. The algorithm can produce the predicate path graph by scanning database only once and dig out the frequent pattern based on the frequent predicate path graph. It avoids the shortcoming of scanning database many times [9].

Usman and Asghar proposed an integrated OLAM architecture that extends the architecture by adding an automation layer for the schema generation. In this paper hierarchical clustering was proposed and three types of schemas namely star, snowflake and galaxy were automatically generated. The enhanced OLAP and data mining system were integrated to achieve the higher degree of development [14]. *Uguz* proposed two-stage feature selection and feature extraction to improve the performance of text categorization. In the first stage, each term within the document is ranked based on their importance for classification. In the second stage, generic algorithm (GA) and principal component analysis (PCA) feature selection and feature extraction methods was applied separately to the terms which are ranked in decreasing order of importance and a dimension reduction was carried out. The experiments were conducted using the k-nearest neighbour (KNN) and decision tree algorithm [15].

Manda et al proposed an approach called Multi-ontology data mining at All Levels (MOAL) which uses the structure and relationships of the Gene Ontology (GO) to mine multi-ontology multi-level association rules. In this paper, two interesting measures was introduced i.e. Multi-ontology Support (MOSupport) and Multi-ontology Confidence (MOConfidence) customized to evaluate multi-ontology multi-level association rules [18].

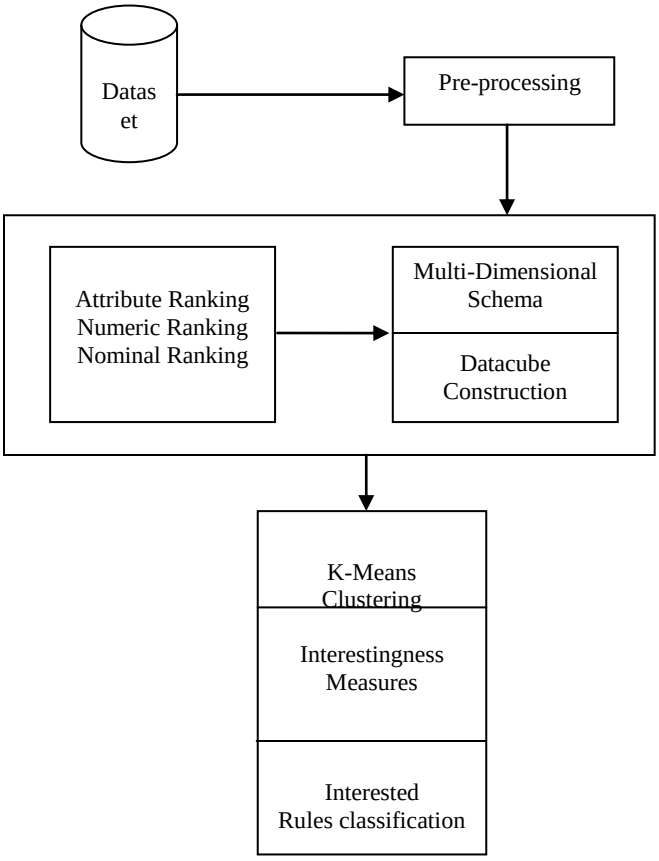
Qodmanan et al proposed a method based on genetic algorithm without consider the minimum support and confidence into account. FP tree algorithm was applied to improve the algorithm efficiency. This method extracts the best rules that have best correlation between support and confidence [19]. An algorithm was proposed to mine fuzzy association rules from uncertain data which was represented by possibility distribution [20].

DCAR

Data cube Association Rule is termed on DCAR. Data Mining and statistical techniques are occupied with PCA to rank the facts and dimensions. Attributes are ranked based on nominal and numerical attributes. Information gain is utilised to rank the numerical attributes. Highly ranked dimensions and facts discovers interesting information nested in multi-dimensional cubes. Datacube is constructed using highest ranked dimensions and facts. Present in multi-dimensional scheme. It mines the association rules based on the importance among the rules form the schema.[22]

III. MODIFIED APPROACH FOR CLASSIFYING MULTI-DIMENSIONAL DATA CUBE THROUGH ASSOCIATION RULE MINING [CMDC]

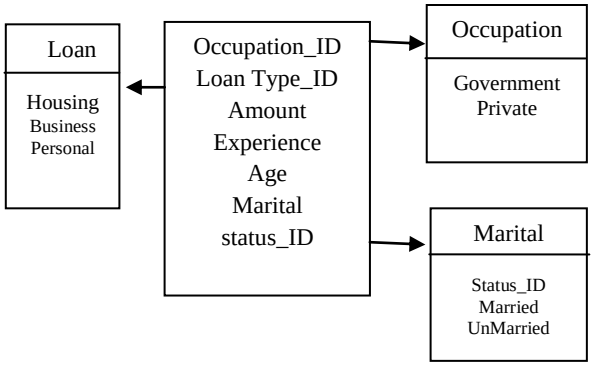
This section presents an overview of Multi-dimensional schema formation which suggests the discovery of Association Rules. Exploring knowledge discovery process by integrating machine learning and statistical scheme. Highly ranked dimensions and fact results in the discovery of interesting information nested in multi-dimensional cubes. The following figure shows the methodology for classifying Multi-Dimensional Data cube. The real world bank loan dataset is used to diverse association rules. The dataset is initially perform pre-processing stages to attain the highest quality of the dataset. It removes the unwanted data in the dataset. The proposed system does not depend on hierarchical structure. It applies k-means clustering techniques for grouping the data which provides better results than other existing methods.



Proposed Algorithm

Input : Bank dataset
Output: Interested Rules Classifications
Step 1: Attributes Ranking [PCA Methods]
Step 2: Multi-Dimensional Scaling
Step 3: Data cube Construction
Step 4: Association Rule Formation
Step 5: Grouping Integrated Rules based on K-Means Clustering techniques.

Multi-dimensional schema is composed of a set of data cubes. The main purpose of a schema is to impose the high level polices. Before proceeding with Multi-dimensional schema, semantic relationship is calculated for each nominal attributes which can be easily visualized by parallel co-ordinates. After receiving the ranked list of nominal and numerical attributes, Multi-dimensional galaxy schema is formed. Here fact tables share the dimensional values. Physical structure is formed based on SQL Queries. For example, Multi-dimensional schema is generated for bank loan dataset is shown in figure.



The above multi-dimensional schema contains all the dimensional and facts in the bank Loan dataset. The schema is used to construct the informative datacubes.

Interestingness measures:

Interestingness of the generated rule is evaluated with the help of alternative evaluation measure. It is used to capture the usefulness of a rule. Importance of rule can be classified as support, confidence and weight formula.

K-Means Clustering

Similar objects are grouped together which automatically reduces the time complexity

Interested Rule classification:

Discovered Rules are classified into highly interested, medium and low interested rules. K-means clustering method is applied to classify Interesting when in this approach.

To classify the rules, min-threshold limit and maximum threshold limit and maximum threshold limit is taken as input and rule set is classified based on the limit value. Interested rules are classified as Low(LR), Medium(MR) and High(HR) rule set.

Interested rule classification
Input: Rule set Rs, Min_TL, Max_TL
Output: LR, MR, HR
for each $r_i \in Rs$
get $weight_i$ for r_i
if $weight_i \leq min\_TL$
add r_i into LR
if $weight \geq Min\_TL$ && $weight_i \leq Max\_TL$
add r_i into MR
if $weight_i > Max\_TL$
add r_i into HR
end
The above algorithm categories the given ruleset Rs in to three disjoint sets such as LR, MR and HR.

Highly Interested Rules

Rule Number	Rules	Weight
R1	Occupation=Govt, Loantype=Personal, Marital_Status=Unmarried	0.985
R2	Occupation=Private Loantype=Business Marital_Status=Married	0.895
R3	Occupation=Govt, Loantype=Housing Marital_Status=Unmarried	0.755

Low Interested Rules

R6	Occupation=Govt, Loantype=Personal, Marital_Status=Married	0.215
R7	Occupation=Private, Loantype=Business, Marital_Status=Married	0.195

Medium Interested Rules

R9	Occupation=Govt, Loantype=Bussiness, Marital_Status=Unmarried	0.668
R5	Occupation=Private, Loantype=Personal, Marital_Status=Married	0.545

IV. PERFORMANCE ANALYSIS

The proposed methodology is evaluated to validate the prediction accuracy and classification of association rules. The experimental dataset is generated by own. Initially , the dataset selected and undergoes pre-processing steps to filter unwanted data present in the particular dataset. The loan applicants are classified based on loantype, occupation and marital status. Experimental results of the proposed method is validated based on the metrics such as prediction accuracy, objective diverse measure and time consumption.

An obvious way to assess the quality of learned model is to observe the prediction accuracy given by the model. Figure shows that the prediction accuracy of the rules generated thro' the use of Multi-dimensional schema which is higher than the existing DCAR.

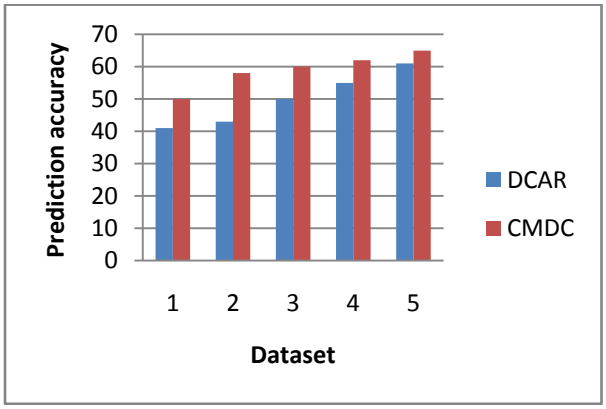


Figure1 Prediction Accuracy

Computing time for proposed algorithm takes lesser time then the existing DCAR method.

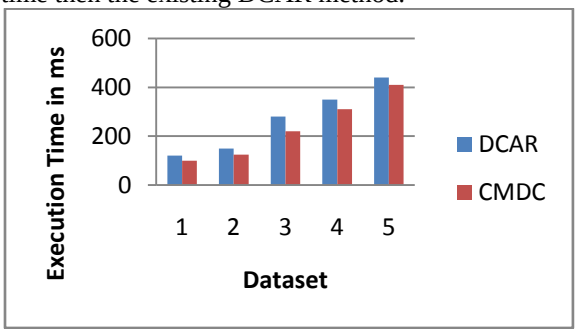


Figure 2 Time Comparison

V.CONCLUSION

This paper proposed modified approach to generate a multi-dimensional schema for classifying interested rules. It filters the non-information dimension and facts from the initial database using principal component Analysis. The prediction accuracy of generated interesting rules are higher than the rules generated by existing method. And also computing time is lesser then existing system.

REFERENCES

[1] M. Usman, R. Pears, and A. Fong, "Discovering diverse association rules from multidimensional schema," 2013.
[2] R. Pears, M. Usman, and A. Fong, "Data guided approach to generate multi-dimensional schema for targeted knowledge discovery," 2012.
[3] G. Liu, H. Jiang, R. Geng, and H. Li, "Application of multidimensional association rules in personal financial services,"

in Computer Design and Applications (ICCDA), 2010 International Conference on, 2010, pp. V5- 500-V5-503.
[4] W.-Y. Chiang, "To mine association rules of customer values via a data mining procedure with improved model: An empirical case study," Expert Systems with Applications, vol. 38, pp. 1716-1722, 2011.
[5] M. A. Domingues and S. O. Rezende, "Using taxonomies to facilitate the analysis of the association rules," arXiv preprint arXiv:1112.1734, 2011.
[6] T. Herawan and M. M. Deris, "A soft set approach for association rules mining," Knowledge-Based Systems, vol. 24, pp. 186-195, 2011.
[7] V. Kumar and A. Chadha, "Mining Association Rules in Student's Assessment Data," International Journal of Computer Science Issues, vol. 9, pp. 211-216, 2012.
[8] C. Romero, J. R. Romero, J. M. Luna, and S. Ventura, "Mining Rare Association Rules from e-Learning Data," in EDM, 2010, pp. 171-180.
[9] H. Zhu and Q. Li, "An Algorithm Based on Predicate Path Graph for Mining Multidimensional Association Rules," in Proceedings of the 2012 International Conference on Information Technology and Software Engineering, 2013, pp. 783-791.
[10] C.-A. Wu, W.-Y. Lin, C.-L. Jiang, and C.-C. Wu, "Toward intelligent data warehouse mining: An ontology-integrated approach for multi-dimensional association mining," Expert Systems with Applications, vol. 38, pp. 11011-11023, 2011.
[11] J. K. Chiang and H. Sheng-Yin, "Multidimensional data mining for healthcare service portfolio management," in Computer Medical
CiiT International Journal of Data Mining and Knowledge Engineering, Vol 6, No 07, August 2014
Applications (ICCMA), 2013 International Conference on, 2013, pp. 1-8.
[12] P. Allard, S. Ferré, and O. Ridoux, "Discovering Functional Dependencies and Association Rules by Navigating in a Lattice of OLAP Views," in CLA, 2010, pp. 199-210.
[13] W. Moudani, M. Hussein, M. Moukhtar, and F. Mora-Camino, "An intelligent approach to improve the performance of a data warehouse cache based on association rules," Journal of Information and Optimization Sciences, vol. 33, pp. 601-621, 2012.
[14] M. Usman and S. Asghar, "An Architecture for Integrated Online Analytical Mining," Journal of Emerging Technologies in Web Intelligence, vol. 3, pp. 74-99, 2011.
[15] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," Knowledge-Based Systems, vol. 24, pp. 1024-1032, 2011.
[16] W. Abdelbaki, S. B. Yahia, and R. B. Messaoud, "NAP-SC: A Neural Approach for Prediction over Sparse Cubes," in Advanced Data Mining and Applications, ed: Springer, 2012, pp. 340-352.
[17] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Association rule mining to detect factors which contribute to heart disease in males and females," Expert Systems with Applications, 2012.
[18] P. Manda, F. McCarthy, and S. M. Bridges, "Interestingness measures and strategies for mining multi-ontology multi-level association rules from gene ontology annotations for the discovery of new GO relationships," Journal of biomedical informatics, 2013.
[19] H. R. Qodmanan, M. Nasiri, and B. Minaei-Bidgoli, "Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence," Expert Systems with applications, vol. 38, pp. 288-298, 2011.
[20] C.-H. Weng and Y.-L. Chen, "Mining fuzzy association rules from uncertain data," Knowledge and Information Systems, vol. 23, pp. 129-152, 2010.
[21] N. Zbidi, S. Faiz, and M. Limam, "On mining summaries by objective measures of interestingness," Machine learning, vol. 62, pp. 175-198, 2006.
[22] K.Kala "DCAR: A Novel Approach for Datacubes Association Rule Algorithm in Multidimensional Schema" CiiT International Journal of Data Mining and Knowledge Engineering, Vol 6, No 07, August 2014