www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume – 5 Issue -03 March, 2016 Page No. 15976-15979

Overview of Classification and Risk Evaluation in Multi-dimensional Datasets

Dr. K.Kavitha
Assistant Professor, Department of Computer Science
Mother Teresa Women's University, Kodaikanal
kavitha.urc@gmail.com

Abstract: ***Data mining is the task of discovering useful and interested patterns from the huge amount of the data where the data can be stored in databases, data warehouses and other information repositories. Data mining comprises an integration of techniques from various disciplines such as data visualization, database technology, information retrieval, high performance computing, machine learning and pattern recognition, etc. The classification of multi-dimensional data is one of the major challenges in data mining and data warehousing. In a classification problem, each object is defined by its attribute values in multidimensional space. Some of the existing systems consider the data analysis might identify the set of candidate data cubes for exploratory analysis based on domain knowledge. Unfortunately, conditions occurred for such assumptions are not valid and these include high dimensional databases, which are difficult or impossible to pre-calculate the dimensions and cubes. Some proposed system is formulated automatically find out the dimensions and cubes, which holds the informative and interesting data. In high dimensional datasets, the data analysis procedures need to be integrated with each other. Based on the information theoretic measures like Entropy is used to filter out the irrelevant data from the dataset in order to formulate a more compact, manageable and useful schema.***

Keyword: Data Mining, Risk Prediction, Association Rule, Information Gain

1. 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.

Data mining techniques uses some key ideas for data classification and prediction. Clustering techniques is used to place data items in to similar groups without prior knowledge of group definitions. Clustering provides efficient decision making by grouping large voluminous datasets in bank. Risk assessment is an important task of bank, as the increase and decrease of credit limits in bank depends largely to evaluate the risk properly. The key problem consists of identifying good and bad customer's status those who applied for loan. An improvised risk evaluation of Multi-dimensional Risk prediction clustering Algorithm is proposed to determine the good and bad loan applicants whether they are applicable or not.

In order to increase the accuracy of risk, risk assessment is performed in primary and secondary levels. Hence for avoiding Redundancy, Association Rule is integrated. This method allows for finding the risk percentage to determine whether loan can be sanctioned to a customer or not.

2. OBJECTIVES

The major objectives of this research work is

DOI: 10.18535/ijecs/v5i3.16

- To Presents the multidimensional data based on associative clustering algorithm.
- To evaluate the risk present on multidimensional data based on the risk prediction clustering algorithm.
- To provide an automated support in multidimensional schema, and identifying the cubes of interest from the multidimensional data.
- To enable a knowledge discovery from the various association rules from large multidimensional cube structure.
- To categorize the interested rules into highly interested, medium interested and low interested rules.

3. Multi-Dimensional Data Cube Construction Models

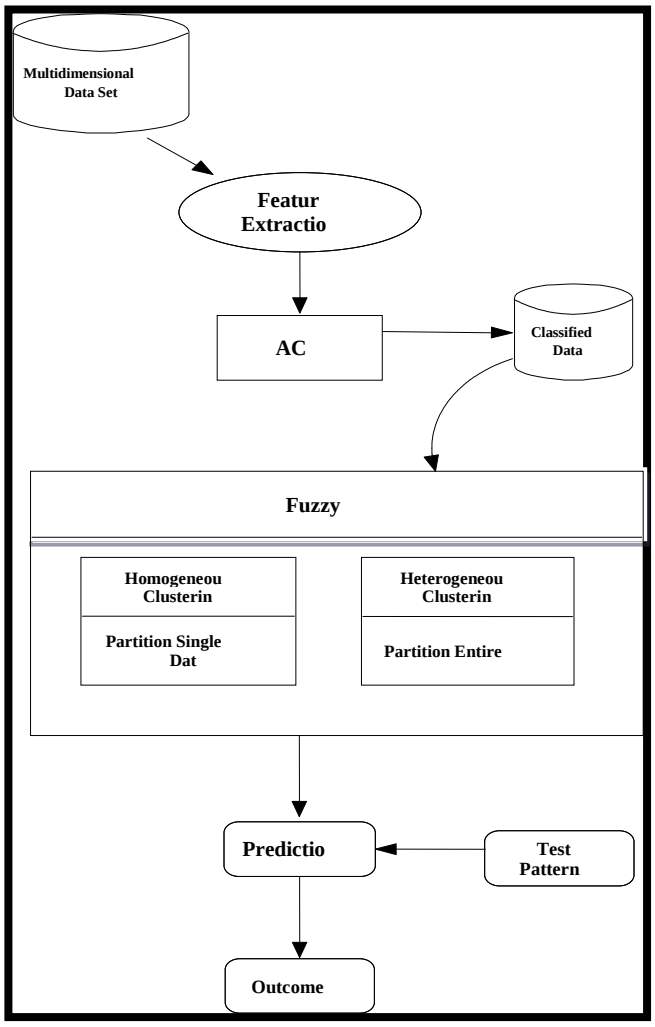
3.1. ACM Methodology

To classify the multidimensional data based on associative clustering model .

Multidimensional data classification is an important and challenging problem in data warehousing and large database environment. Contribution of Naive Bayes classifier in proposed ACM model is proving to be versatile and robust in multidimensional data classification. Proposed ACM model effectively retrieves the predicted information from multidimensional dataset. PCA is exploited to extract the features, and then multidimensional data is loaded to perform the task.

Description of multidimensional data set:

The data set used in this work is multidimensional data set. The data set contains different attributes which include Age, Job, Marital status, Education, Default, Balance, Housing loan, Personal loan, Contact, Day, Month, Duration, Campaign, P days, Previous, P outcome, Term deposit.



ACM framework

The given data set is usually divided into training and prediction phase.During training phase, feature extractor is used to convert each unseen input value to a feature set.The basic information about each input is captured by these feature set and is classified. Moreover, combination of both label and feature set are fed into the knowledge discovery technique (Machine learning algorithm) to generate model. The clustered data are then fed into the prediction phase to generate predicted labels. A model is build using training data set and validates the model with prediction set. In the proposed ACM framework, the attributes are initialized. Principal Component Analysis (PCA) is used to reduce the features used to represent the multidimensional data and it provides fast classification, simple representation and reduction in memory. The

DOI: 10.18535/ijecs/v5i3.16

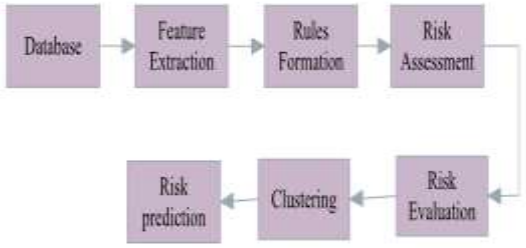
given data set is usually divided into training and prediction phase and the information about each input is captured by feature set and is classified. This classifier model segments the data. In clustering model, the classified data are grouped or clustered and the clustered data are then fed into the prediction phase to generate predicted labels.

A model is build using training data set and validates the model with prediction set. From the given collection of multidimensional data set (training set), each record contains a set of features, one of the feature is a class.

Test data is used to determine the accuracy of the model. The Naive Bayes classifier is used to classify the multidimensional data. The learned patterns cld1, cld2, cld3 are applied to this test set t and the resulting output is compared to the desired output.If the learned patterns do meet the desired standards, then the final step is to interpret the learned patterns and turn them into knowledge p1, p2.The outcome of this test pattern produce information retrieved from multidimensional data set.

3.2. ERPCA: a novel approach for risk evaluation of multidimensional data based on efficient risk prediction clustering algorithm

To evaluate the risk evaluation of multidimensional data based on risk prediction clustering algorithm. Credit scoring is defined as a statistical method that is used to predict the probability that a loan applicant will default or become delinquent.. Credit scoring helps to increase the speed and consistency of the loan application process and allows the automation of the lending process. Risk assessment is one of the existing problems in the bank sector. The decision for the credit sanction to a customer should be evaluated properly so that, it may not lead to loss for the Bank. The proposed method (ERPCA) aids the banking sector to make the evaluation for loan sanction in an enhanced manner. Rules are formed for each loan type like (personal loan, bike loan, car loan, house loan, business loan) [13].



Overall Flow of ERPCA

Clustering algorithm

Associative clustering algorithm (ERPCA) is used to mine the clusters from massive and high dimensional numerical databases. A group of data elements can belong to more than one cluster, which is associated with each element is a set of membership levels. Using ERPCA algorithm, three vectors can be taken into consideration. The centroid and coefficient of classified data is computed and the obtained result is compared with three initialized vectors.

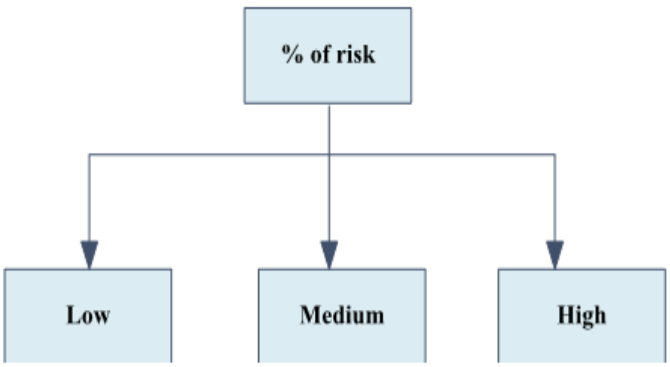
The variables L1, L2, M1, M2, H quoted in this algorithm takes the value of 0 and 25 for low, 26 and 50 for medium and greater than 50 for high. Based on these three vectors, the data are clustered.

Risk Prediction

For the loan sanction, a threshold value of 35% of risk is set. If the % of risk for a customer is greater than 35%, the application is rejected, else loan is sanctioned.

Loan approval list and Loan rejection list are classified using this threshold value.

Then the loan approved customers and loan rejected customers are clustered separately for efficient retrieval.

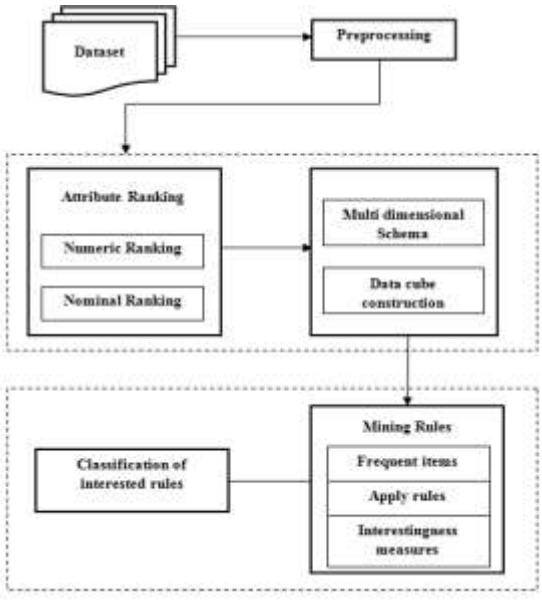


Risk levels

3.3. DCAR: Data cubes association rules algorithm in multi dimensional schema

To generate a multidimensional schema using IRCA algorithm for the classification of the interested rules.The attributes are categorized as nominal and numerical attributes. The numeric attributes are ranked based on Principal Component Analysis (PCA) method; the nominal attributes are ranked based on Information Gain. Multidimensional scaling is applied to find the semantic relationships among the values for each nominal attribute.The informative data cubes are constructed based on the highly ranked facts and dimensions.

The interested rules are classified based on IRCA algorithm.The proposed algorithm categorizes the interested rules into highly interested, medium interested and low interested rules[12].



Methodology for discovering the association rules

The real World Bank loan dataset is used to diverse the association rules. The dataset is initially preprocessed to attain the highest quality of the dataset. Preprocessing step is used to remove the unwanted data in the dataset.This method does not depend on any hierarchical structure and does not performs clustering.

4. CONCLUSION

This paper analysed the papers related with Multi-dimensional data Cubes and highlighted the benefits and limitations in this area.In this analysis, ACM is used for efficient classification of multi dimensional data and to predict the outcome accurately.Risk assessment is a very crucial task in banking industry and ERPCA is used for risk evaluation. In ERPCA, mass volume of customer data are engendered and risk assessment plus evaluation is done based on the Data mining technique.The attributes are selected using Information gain theory.Clustering algorithm is used to classify the risk levels as low, medium and high, based on the percentage of risk values obtained. An IRCA approach used to generate a multidimensional schema and an IRCA algorithm for the classification of the interested rules was introduced. A multidimensional schema was formulated completely with the informative datacubes.

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
[12] 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
[13] K Kala E.Ramaraj ERPCA: A Novel Approach for Risk Evaluation of Multidimensional Risk Prediction Clustering Algorithm", International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397 Vol. 5 No. 10 Oct 2013