International Journal of Advance Research in Science and Engineering
Vol. No. 8, Issue No. 02 , February 2019
www.ijarse.com
IJARSE
ISSN 2319 - 8354

AN INVESTIGATION OF VARIOUS DATA MINING CLUSTER ANALYSIS BASED ON DISCRIMINATION INTERRUPTION IN TEMPORAL DATA MINING.

G Shyama Chandra Prasad

Associate Prof, Matrusri Engineering College, Hyderabad,

ABSTRACT:

Clustering means keeping similar objects together. Document clustering is an extension of clustering, which is related to keeping similar text documents together. Document clustering plays a vital role in development of search engines, where a group of document is required to listed as a result of query in minimum response time. This paper elaborates the concept of document cum text clustering.Extracting useful information hidden in large collection of data is known as data mining. Discrimination is a very important issue when considering the legal and ethical aspects of data mining. Discrimination mean unfairly treating people on the basis of their cast, religion , gender etc. Due to Rapid development in Data mining, Evolution of Temporal data mining is newly evolving research area in vast field of Data Mining.Temporal Data Mining having applications in-following fields such as bio-medicine, geographical data processing, financial data forecasting and Internet site usage monitoring. Temporal data mining deals is a process of extracting of knowledge information from temporal data, where the definition of Knowledge depends on the user application. The most common form of temporal data is time series data, which consist of real values sampled at regular time intervals.

Keywords: ***Data Mining, Utility, Sensitive data item, Temporal Data mining , Discrimination, Privacy Preserving, Decision Tree, Rules.***

I. INTRODUCTION

Data mining and knowledge discovery in databases are two new research areas that investigate the automatic extraction of previously unknown patterns from large amounts of data. Data mining involves the extraction of implicit previously unknown and potentially useful knowledge from large databases.

Data mining extracts novel and useful knowledge from data and has become an effective analysis and decision means in corporation. Results of Data Mining Include:

- Forecasting what may happen in the future
- Classifying people or things into groups by recognizing patterns
- Clustering people or things into groups based on their attributes
- Associating what events are likely to occur together
- Sequencing what events are likely to lead to later events

1.1 Temporal data mining:

Temporal Data Mining (TDM) is defined as the activity of looking for interesting correlations or patterns in large temporal data sets. TDM has evolved from data mining and was highly influenced by the areas of temporal databases and temporal reasoning.

Clustering is a partition of data into groups of related objects. Each set, called cluster, consists of objects which are similar to each other and dissimilar to the item of other groups. In other language, the principle of a high-quality document clustering approach is to decrease intra-cluster distances between documents. It is shown below in figure 2. In clustering is the allocation and the nature of information that will conclude cluster membership, in conflict to the classification where the classifier learn the association between objects and classes from a so set, i.e. a set of documents properly label by hand, and then replicates the learn performance on unlabeled data

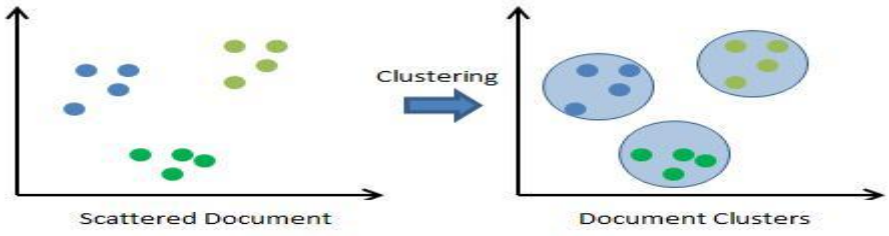

Figure 2: Document Clustering [4]

The document clustering framework is shown below in figure 3. Input are text documents. Then key words are identified in these documents. Then similarity is measured in these documents. Generally Euclidian distance is used as similarity measure.

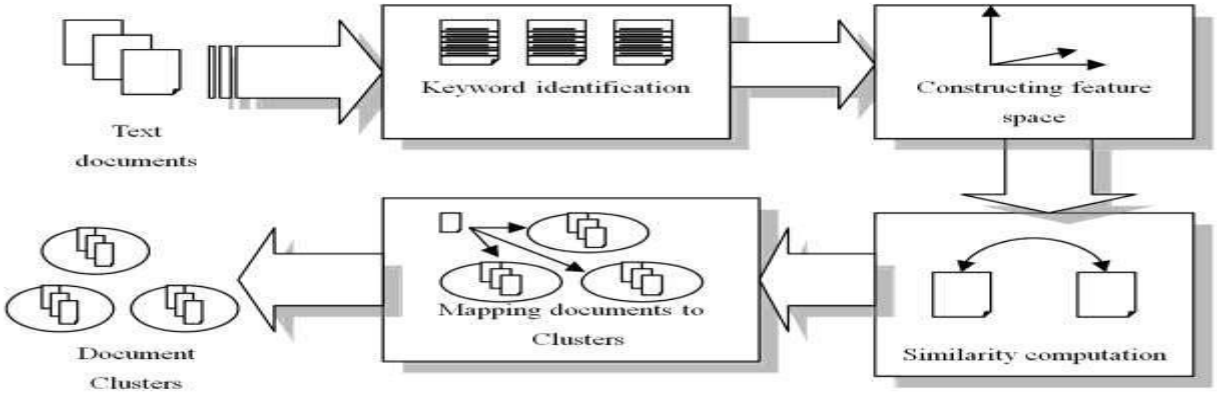

Figure 3: Document Clustering Framework

2.2 Preferential Sampling

Introduced the idea of Classification with No Discrimination (CND). We propose a new solution to the CND problem by we introduce a Preferential Sampling (PS) scheme to make the dataset bias free. Instead, PS changes the distribution of different data objects for a given data to make it discrimination free. To identify the borderline objects , PS starts by learning a ranker on the training data. PS uses this ranker to class the data objects of DP and PP in ascending order, and the objects of DN and PN in descending order both with respect to the positive class probability. Such understanding of data objects makes sure that the higher the rank an element occupies, the closer it is to the borderline.PS starts from the original training dataset and iteratively duplicates and removes objects in the following way Decreasing the size of a group is always done by removing the data objects closest to the borderline. Increasing the sample size is done by duplication of the data object closest to the borderline.

PS works in the following steps:

- Divide the data objects into the four groups, DP,DN, PP, and PN.
- Any ranking algorithm may be used for calculating the class probability of each data tuple. This ranking will be used to identify the borderline data objects.
- Calculate the expected size for each group to make the dataset bias free.

(iv)Finally apply sampling with replacement to increase the size of DP and PN. And decrease the size of DN and PP.

2.2.1 Result

Classification with No Discrimination by Preferential Sampling is an excellent solution to the discrimination problem. It gives promising results with both stable and unstable classifiers give more accurate results but do not reduce the discrimination.

2.2.2 Drawbacks

Low data utility rate and minimum discrimination removal. This PS is also not applicable for Indirect discrimination.

2.3 Decision Tree Learning

This approach in which the non-discriminatory constraint is pushed deeply into a decision tree learner by changing its splitting criterion and pruning strategy by using a novel leaf relabeling approach. We propose the following two techniques for incorporating discrimination awareness into the decision tree construction process:

Dependency-Aware Tree Construction: When evaluating the splitting criterion for a tree node, not only its contribution to the accuracy, but also the level of dependency caused by this split is evaluated.

Leaf Relabeling: Normally, in a decision tree, the label of a leaf is determined by the majority class of the tuples that belong to this node in t he training set. In leaf relabeling we change the label of selected leaves in such a way that dependency is lowered with a minimal loss in accuracy.

2.3.1 Result

This method gives high accuracy and low discrimination scores when applied to non-discriminatory test data. In this scenario, our methods are the best choice, even if we are only concerned with accuracy. The enrichment in

discrimination reduction with the relabeling method is very satisfying. The relabeling methods out-perform the baseline in almost all cases. As such it is reasonable to say that the straightforward solution is not satisfactory and the use of dedicated discrimination-aware techniques is justified.

2.3.2 Drawbacks

The result of this approach has mostly similar to the Naïve Bayesian Approach and it only concerned with accuracy. Discrimination removal is very low using relabeling method.

2.4 Indirect Discrimination Prevention

This Method regarding discrimination prevention is considering indirect discrimination other than direct discrimination and another challenge is to find an optimal trade-off between anti-discrimination and usefulness of the training data. The main contributions of this method are as follows: (1) a new pre-processing method for indirect discrimination prevention based on data transformation that can consider several discriminatory attributes and their combinations (2) some measures for evaluating the proposed method in terms of its success in discrimination prevention and its impact on data quality. This solution is based on the fact that the dataset of decision rules would be free of indirect discrimination if it contained no redlining rule.

Data Transformation Method for Indirect Discrimination:

Rule Protection

The indirect discriminatory measure to convert redlining rules into non-redlining rules, we should enforce the following inequality for each redlining rule r: D, B→C $_{in}$

$elb\ (\gamma, \delta) < \alpha$

In order to implement this data trans-formation method for indirect discrimination prevention, we simulate the availability of a large set of background rules under the assumption that the dataset contains the discriminatory items. The utility measures of indirect discrimination is same as the above preprocessing approach based on the redlining rule dataset RR

2.4.1 Result

The values of DDP and DPD achieves a high degree of indirect discrimination prevention in different cases. In addition, the values of MC and GC demonstrate that this proposed solution incurs little information loss, especially when α is not too small. By decreasing the value of α,the amount of redlining rules is increased, which causes further data transformation to be done, there by increasing MC and GC.

2.4.2 Drawbacks

The execution time of this algorithm increases linearly with the number of redlining rules and α-discriminatory rules. This method is only deal with indirect discrimination and it cannot measure the direct discriminatory items.

2.5 Direct and Indirect Discrimination Prevention Method

This new technique applicable for direct or indirect discrimination prevention individually or both at the same time and effective at removing direct and/or indirect discrimination biases in the original data set while preserving data quality. This method can be described in terms of two phases:

Discrimination measurement- Direct and indirect discrimination discovery includes identifying α discriminatory rules and redlining rules.

(i) Based on predetermined discriminatory items in DB, frequent classification rules in FR are divided in two groups: PD and PND rules.

(ii) Direct discrimination is measured by identifying α- discriminatory rules among the PD rules using a direct discrimination measure and a discriminatory threshold (α).

(iii) Indirect discrimination is measured by identifying redlining rules among the PND rules combined with background knowledge, using an indirect discriminatory measure (elb), and a discriminatory threshold (α).

Data transformation- Transform the original data DB in such a way to remove direct and/or indirect discriminatory biases, with minimum impact on the data and on legitimate decision rules, so that no unfair decision rule can be mined from the transformed data.

2.5.1 Transformation Method

The key problem of transforming data with minimum information loss to prevent at the same time both direct and indirect discrimination. We will give a pre-processing solution to simultaneous direct and indirect discrimination prevention. There are two transformation method used in both direct and indirect discrimination removal.

(i) ***Direct Rule Production*** **-** In order to convert each α-discriminatory rule into a α-protective rule, based on the direct discriminatory measure. *elift (r) <α*

(ii) ***Indirect Rule Protection*** **-** In order to turn a redlining rule into an non-redlining rule, based on the indirect discriminatory measure we should enforce the following inequality for each redlining ruler: D,B Cin RR: *elb (γ, δ) < α* These two data transformation method for used simultaneous direct and indirect discrimination prevention.

2.5.2 Utility Measures

These techniques should be evaluated based on two aspects

- To measure the success of the method in removing all evidence of direct and/or indirect discrimination from the original data set.
- To measure the impact of the method in terms of information loss

2.5.3 Drawbacks

The main drawbacks of this method contain Low privacy assurance and Limited utility ratio of data. The association of privacy is not analysed from the transformed dataset.

Direct and Indirect Discrimination Prevention Method

This new technique applicable for direct or indirect discrimination prevention individually or both at the same time and effective at removing direct and/or indirect discrimination biases in the original data set while preserving data quality. This method can be described in terms of two phases:

Discrimination measurement- Direct and indirect discrimination discovery includes identifying α discriminatory rules and redlining rules.

(iii) Based on predetermined discriminatory items in DB, frequent classification rules in FR are divided in two groups: PD and PND rules.

(iv) Direct discrimination is measured by identifying α- discriminatory rules among the PD rules using a direct discrimination measure and a discriminatory threshold (α).

(iv) Indirect discrimination is measured by identifying redlining rules among the PND rules combined with background knowledge, using an indirect discriminatory measure (elb), and a discriminatory threshold (α).

Data transformation- Transform the original data DB in such a way to remove direct and/or indirect discriminatory biases, with minimum impact on the data and on legitimate decision rules, so that no unfair decision rule can be mined from the transformed data.

2.5.1 Transformation Method

The key problem of transforming data with minimum information loss to prevent at the same time both direct and indirect discrimination. We will give a pre-processing solution to simultaneous direct and indirect discrimination prevention. There are two transformation method used in both direct and indirect discrimination removal.

(iii) ***Indirect Rule Protection*** - In order to turn a redlining rule into an non-redlining rule, based on the indirect discriminatory measure we should enforce the following inequality for each redlining ruler: D,B Cin RR: $elb(\gamma, \delta) < \alpha$ These two data transformation method for used simultaneous direct and indirect discrimination

prevention.

2.5.2 Utility Measures

These techniques should be evaluated based on two aspects

- To measure the success of the method in removing all evidence of direct and/or indirect discrimination from the original data set.
- To measure the impact of the method in terms of information loss

2.5.3 Drawbacks

The main drawbacks of this method contain Low privacy assurance and Limited utility ratio of data. The association of privacy is not analyse from the transformed dataset.

III. PROPOSED SOLUTION

The main negative impacts of data mining is discrimination and privacy. The privacy is connection with current privacy models, like differential privacy. It will provide the high privacy rate. This method is integrated with the previous existing method of direct and indirect discrimination prevention mechanism and to find synergies between rule hiding for privacy preserving data mining and association rule hiding for discrimination removal. Rule privacy is optimized with rule generalization mechanism. These methods provide the competent outcome of removing the discrimination with high privacy rate.

IV. SIMULATION ANALYSIS AND RESULT

To measure the effectiveness of PPTDM algorithm, the experiments were conducted on synthetic dataset and compared by the hiding failure. All experiments were performed on a Dell workstation with 3.40 GHz Intel Pentium 4 processor and 2 GB of main memory running the Windows XP professional and simulation performed on WEKA simulation tool and use netbeans7.3 as IDE(Integrated development environment) for deploy of project.

WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code.

WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes[22].

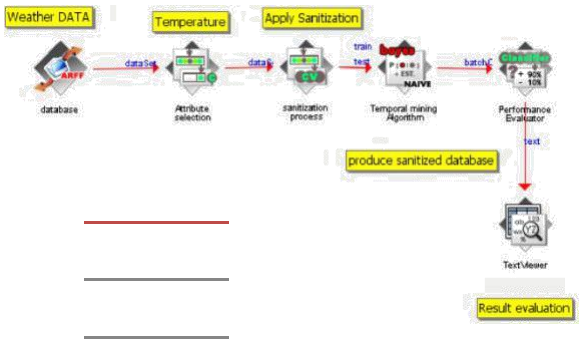

Fig 4.1 knowledge flow of Temporal mining process

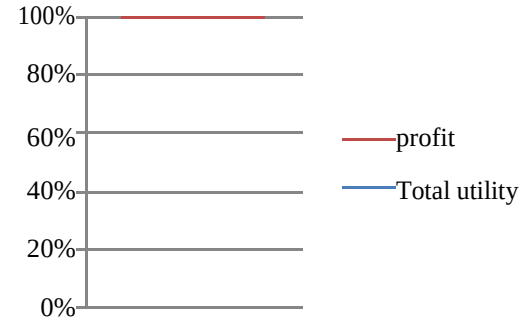

Fig 3.2.2 shows total utility of various data items

IV. CONCLUSION

In this paper, we have completed a wide overview of the distinctive methodologies for discrimination prevention for data mining. We discussed the issues and limitation of the recent state of the approaches. Based on the same issues, we study an approach that uses transformation method. This approach helps to prevent direct discrimination and indirect discrimination.The focus is on Document Clustering which is very recent technology, we investigated many existing algorithms. As clustering plays a very vital role in various applications, many researches are still being done. The upcoming innovations are mainly due to the properties and the characteristics of existing methods. This paper presents an introduction to the present document clustering concept along with the methods used for document clustering. A critical review of existing work done by authors on document clustering in recent time is also presented in this paper.

REFERENCES

[1] D. Brillinger, editor. Time Series: Data Analysis and Theory. Holt, Rinehart and Winston, New York, 1975.

[2] P. Cheeseman and J. Stutz. Bayesian classification (AUTOCLASS): Theory and results. In U. M.

Fayyad,G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy,editors, Advances in Knowledge Discovery and DataMining. AAAI Press / MIT Press, 1995.

[3] T. Fulton, S. Salzberg, S. Kasif, and D. Waltz. Local induction of decision trees: Towards interactive data mining. In Simoudis et al. [21], page 14.

[4] B. R. Gaines and P. Compton. Induction of metaknowledge about knowledge discovery. IEEE Trans. OnKnowledge And Data Engineering, 5:990–992, 1993.

[5] C. Glymour, D. Madigan, D. Pregibon, and P. Smyth. Statistical inference and data mining. Communicationsof the ACM, 39(11):35–41, Nov. 1996.

[6] F. H. Grupe and M. M. Owrang. Data-base mining -discovering new knowledge and competitive advantage. Information Systems Management, 12:26–31, 1995.

[7] J. Han, Y. Cai, and N. Cercone. Knowledge discovery in databases: An attribute-oriented approach. In Proceedings of the 18th VLDB Conference, pages 547–559,Vancouver, British Columbia, Canada, Aug. 1992.

[8] J. W. Han, Y. D. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. Ieee Trans. On Knowledge And Data Engineering, 5:29– 40, Feburary 1993.

[9] J. W. Han, Y. Yin, and G.Dong. Efficient mining of partial periodic patterns in time series database. IEEE Trans. On Knowledge And Data Engineering, 1998.

[10] D. Heckerman, H. Mannila, D. Pregibon, and R. Uthurusamy, editors. Learning bayesian networks: the combineation of knowledge and statistical data. AAAI Press, 1994.

[11] D. Heckerman, H. Mannila, D. Pregibon, and R. Uthurusamy, editors. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97). AAAI Press, 1997.

[12] E. Keogh and P. Smyth. A probabilistic approach to fast pattern matching in time series databases. Page 126.

[13] A. Ketterlin. Clustering sequences of complex objects. In Heckerman et al. [11], page 215.

[14] C. Li and G. Biswas. Temporal pattern generation usinghidden markov model based unsuperised classifcation. In Proc. of IDA-99, pages 245–256, 1999.

[15] M.J.Zaki. Fast mining of sequential patterns in very large databases. Uni. of Rochester Technical report,1997.

[16] S. a. O.Etzion, editor. Temporal databases: Researchand Practice. Springer-Verlag,LNCS1399, 1998.

[17] B. Padmanabhan and A. Tuzhilin. Pattern discovery in temporal databases: A temporal logic approach. InSimoudis et al. [21], page 351.

[18] P.sprites, C.Glymour, and R.Scheines. Causation, Prediction and Search. Springer-Verlag, 1993.

[19] J. Roddick and M. Spiliopoulou. A survey of temporal knowledge discovery paradigms and methods.IEEE Transactions on Knowledge and Data Engineering, 2002.