AN EFFICIENT CLUSTERING MODEL IN HIGH UTILITY ITEMSET MINING

Dr.R.Kanimozhi,

Head of the department,

Department of BCA,

Idhaya college for women, kumbakonam

ABSTRACT

Mining high utility item sets is considered to be one of the important and challenging problems in the data mining literature. The main objective of Utility Mining is to identify the item sets with highest utilities, by considering profit, quantity, cost or other user preferences. Mining High Utility item sets from a transaction database is to find item sets that have utility above a user-specified threshold. Item set Utility Mining is an extension of Frequent Item set mining, which discovers item sets that occur frequently. In many real-life applications, high-utility item sets consist of rare items.The important consideration in privacy preservation is to provide a proper balance between privacy protection and knowledge discovery. In order to handle these

scenarios carefully, we propose a privacy preserving utility mining method in this paper based on the process, namely, clustering, measure reduction in mining and post-reduction of sensitivity. Rare item sets provide useful information in different decision-making domains such as business transactions, medical, security, fraudulent transactions, retail communities.The proposed approach is designed to handle privacy protection effectively using these three ways. Here, the utility pattern mining algorithm is devised utilizing the data and then, the privacy protection schemes are applied.

1. INTRODUCTION

1.1 Data Mining:

Data Mining (DM) is used to extract the hidden information from the huge volume of data storage. Data Mining plays a key role in a particular field especially in a termed learning analytics or educational data mining to examine for a quality assortment. It seeks to find ways to make helpful use of the enlarging amount of data about learners to understand the process of learning and the social motivational factors encircle learning **[1,2].** The data collected from different applications involve the proper method of extracting knowledge from a large repositories for enhanced decision - making. Knowledge discovery in databases (KDD), often called data mining, aims at a discovery of useful information from a large collections of data **[1,3].**

Nowadays, there's a rapid growing demand of spatial information and spatial information system is well identified. In many areas, large quantities of data are generated and collected, therefore, tremendous spatial data in spatial database management system and spatial data warehouse is used for discovering previously unknown knowledge, and geographical pattern in experimental datasets is frequently investigated. To find useful

information in spatial data, many methods are introduced like association rules, classification, clustering and etc. Classification is particularly important when applied to the analysis of financial, economical, environmental, and demographic phenomena where the data are potentially large, complex **[4].**

Data Mining tasks can be classified into two categories, Descriptive Mining and Predictive Mining. The Descriptive Mining techniques such as Clustering, Association Rule Discovery, Sequential Pattern Discovery, is used to find human-interpretable patterns that describe the data. The Predictive Mining techniques like Classification, Regression, Deviation Detection, use some variables to predict unknown or future values of other variables.

1.2.Association rule mining

It is a popular technique for finding co-occurrences, correlations, frequent patterns, associations among items in a set of transactions or a database. Rules with confidence and support above user-defined thresholds were found. As data continues to grow and its complexity increases, newer data structures and algorithms are being developed to match this development. Association Rule Mining process can be divided into two steps. The first step involves finding all frequent item sets in databases. Once the frequent item sets are found association rules are generated **[6].** Association rule mining is widely used in market-basket analysis. For example, frequent item sets can be found out by analyzing market basket data and then association rules can be generated by predicting the purchase of other items by conditional probability **[1], [2].**

Association rule mining strategies find fascinating associations and relationships among the known collection of information. An association rule is a rule, which involves certain associations with articles or things, for instance, the interrelationship of the information thing as whether they happen all the while with other information thing and how regularly. These rules are registered from the information and, association rules are ascertained with help of likelihood. It has a mentionable measure of viable applications including order, XML mining, spatial information examination, and offer market along with suggestion frameworks.

1.3.Utility Mining:

The traditional ARM approaches consider the utility of the items by its presence in the transaction set. The frequency of item set is not sufficient to reflect the actual utility of an item set. For example, the sales manager may not be interested in frequent item sets that do not generate significant profit. Recently, one of the most challenging data mining tasks is the mining of high utility item sets efficiently. Identification of the item sets with high utilities is called as Utility Mining. The utility can be measured in terms of cost, profit or other expressions of user preferences. For example, a computer system may be more profitable than a telephone in terms of profit. Utility mining model was proposed in **[19]** to define the utility of item set. The utility is a measure of how useful or profitable an item set X is. The utility of an item set X, i.e., u(X), is the sum of the utilities of item set X in all the transactions containing X. An item set X is called a high utility item set if and only if u(X) >= min_utility, where min_utility is a user defined minimum utility threshold [11]. The main objective of high-utility item set mining is to find all those item sets having utility greater or equal to user-defined minimum utility threshold.

2. Literature survey

In the previous section we have introduced the basic concept of Data Mining, Association Rule mining, Utility Mining and Rare Item set Mining. A brief overview of various algorithms, concepts and techniques defined in different research papers have been given in this section. The mining of association rules for finding the relationship between data items in large databases is a well studied technique in data mining field with representative methods like Apriori **[1], [2].** ARM process can be decomposed into two steps. The first step involves finding all frequent item sets in databases. The second step involves generating association rules from frequent item sets.

In **[6], Yao et al** defined the problem of utility mining, a theoretical model called MEU, which finds all item sets in a transaction database with utility values higher than the minimum utility threshold. The mathematical model of utility mining was defined based on utility bound property and the support bound property. This laid the foundation for future utility mining algorithms.

H. **Yao et al** formalized the semantic significance of utility measures in **[11]**. Based on the semantics of applications, the utility-based measures were classified into three categories, namely, item level, transaction level, and cell level. The unified utility function was defined to represent all existing utility-based measures. The transaction utility and the external utility of an itemset was defined and general unified framework was developed to define a unifying view of the utility based measures for itemset mining. The mathematical properties of the utility based measures were identified and analyzed. High utility frequent itemsets contribute the most to a predefined utility, objective function or performance metric**[13]**. For example, From marketing strategy perspective, it is important to identify product combinations that have a significant impact on company's bottom line i.e. having highest revenue generating power**[13]**. An algorithm for frequent item set mining was presented by **Hu et al in [13]** that identify high utility item combinations. In contrast to the traditional association rule and frequent item mining techniques, the goal of the algorithm is to find segments of data, defined through combinations of few items (rules), which satisfy certain conditions as a group and maximize a predefined objective function. The high utility pattern mining problem considered is different from former approaches, as it conducts "rule discovery" with respect to individual attributes as well as with respect to the overall criterion for the mined set, attempting to find groups of such patterns that combined contribute the most to a predefined objective function **[13].**

In the paper [17], H.F. Li proposed two efficient one pass algorithms, MHUI-BIT and MHUI-TID, for mining high utility item sets from data streams within a transaction-sensitive sliding window. Two effective representations of item information and an extended lexicographical tree-based summary data structure were developed to improve the efficiency of mining high utility item sets **[11].**

Liu et al proposed Two-Phase algorithm **[8]** for finding high utility item sets. In the first phase, a model that applies the "transaction-weighted downward closure property" on the search space to expedite the identification of candidates. In the second phase, one extra database scan is performed to identify the high utility item sets.

3.PROBLEM DEFINITION

The huge application of data mining technologies have raised concerns about securing information against unauthorized access is an important goal of database security and privacy. The privacy and security are extra vital essentials when data is distributed. Privacy is a term which is associated with mining task so that we are able to a hide some crucial information which we don't want to disclose to the public. A successful way for prospect data mining study will be the growth of techniques that include privacy concerns. According to that, the privacy preserving data mining has received a bunch of considerate at mutually in research and applications with an important challenge of doing privacy preservation, by providing the proper balance between the privacy protection and knowledge discovery.The primary intention of the proposed technique is to design and develop a technique for privacy protection for the utility patterns in knowledge discovery process. The important consideration in privacy preservation is to provide the proper balance between the privacy protection and knowledge discovery. In the proposed technique, we are considering quality factor for the calculation of utility value apart from the transaction and profit normally used by existing techniques. The quality factor is important as it has a large impact on the future sales of the product. Having higher quality factor will hence lead to better sales. The proposed technique also modifies the reduction formula when compared with existing techniques so as to achieve better results.

3. METHODOLOGY:

3.1.Knowledge Discovery

Data Mining is the core part of the Knowledge Discovery Process. Knowledge Discovery (KD) is an ongoing dynamic process where the data, presented as large repositories, usually complex and heterogeneous in nature, are analyzed and modeled into usable and understandable patterns, often changed into exploratory research ideas for better output understanding. The algorithms that are designed to explore the data sets in Knowledge Discovery procedures and infer a modeling approach of problem-solving to find unidentified patterns are usually based on Data Mining schema. This schema can be used for further studies and can be outlined as a model or as a framework or as an algorithm, for which the specified data sets form the essential elements. The main element in the Knowledge Discovery process is to create the concept from the available data which can then be applied to the existing procedures to get a reasonable knowledge with all the required insights present as patterns. The proposed framework is built by employing two approaches and all two approaches are studied carefully and are applied at correct intervals to give an analytical and systematic background for the research work. The theories employed are Grounded Theory Approach and Transparent Analysis.

3.2. Mining of clusters

Cluster refers to a group of similar kind of objects. Cluster analysis refers to the formation of a group of objects that are very similar to each other but are extremely different from the objects in other clusters.Different data mining algorithms have different approaches to data analysis and therefore present different utility considerations. Therefore, to improve the tradeoff between privacy and utility, 'one size fits all' solution would not be optimal. For example, using a classification metric to guide an anonymization algorithm could provide anonymized data that can reasonably be used for classification tasks. However, to reach better accuracy, a better approach is to adapt a specific data mining algorithm such that it provides a k-anonymous outcome. Similarly, a

framework such as PINQ has a huge advantage such that it makes privacy-preserving data analysis approachable for programmers who have no expertise in privacy research. However, to achieve a good tradeoff between privacy and the accuracy, it is better to develop a differentially-private version of the desired data mining algorithm.

A step by step procedure utilized in this work is given as follows:

Step1: Preprocessing: Identifying the missing values can be done either by eliminating the record or by replacing the missing values by calculating the mean.

Step2: Classification Find the performance of classification algorithms based on its accuracy.

Step3: Clustering

i) Applying K-means Algorithm

K-means is a measurable, unsupervised, non-deterministic, iterative technique for gathering the diverse articles into clusters. The k-means calculation has ended up being to a great degree intense in making clusters in various handy applications in creating regions, for instance, in Bioinformatics, advertise division, PC vision geostatistics, stargazing and cultivation. It is least complex unsupervised learning calculations known for its speed, ease, and convenience.

ii) Applying Fuzzy C-means Algorithm

Fuzzy C-Means is withal named as soft clustering. The FCM calculation endorses a data point to have a place with every one of the groups with the participation in the middle of 0 and 1. If the data point is more proximate to the cluster focus its enrollment will be more towards a specific group. At the end of each emphasis both the inside and enrolment should be refreshed appropriately. FCM alongside Genetic calculation can be adjusted to cause a recommender framework predicated on homogeneous characteristic measures.

Step 4: Prediction

According to the compactness characteristics of data sets of the crop database, the classifiers are trained by the right samples that are chosen through by multi-layer fuzzy C-means clustering, which is the unsupervised learning method. So this method can maintain the accuracy of the classifier.

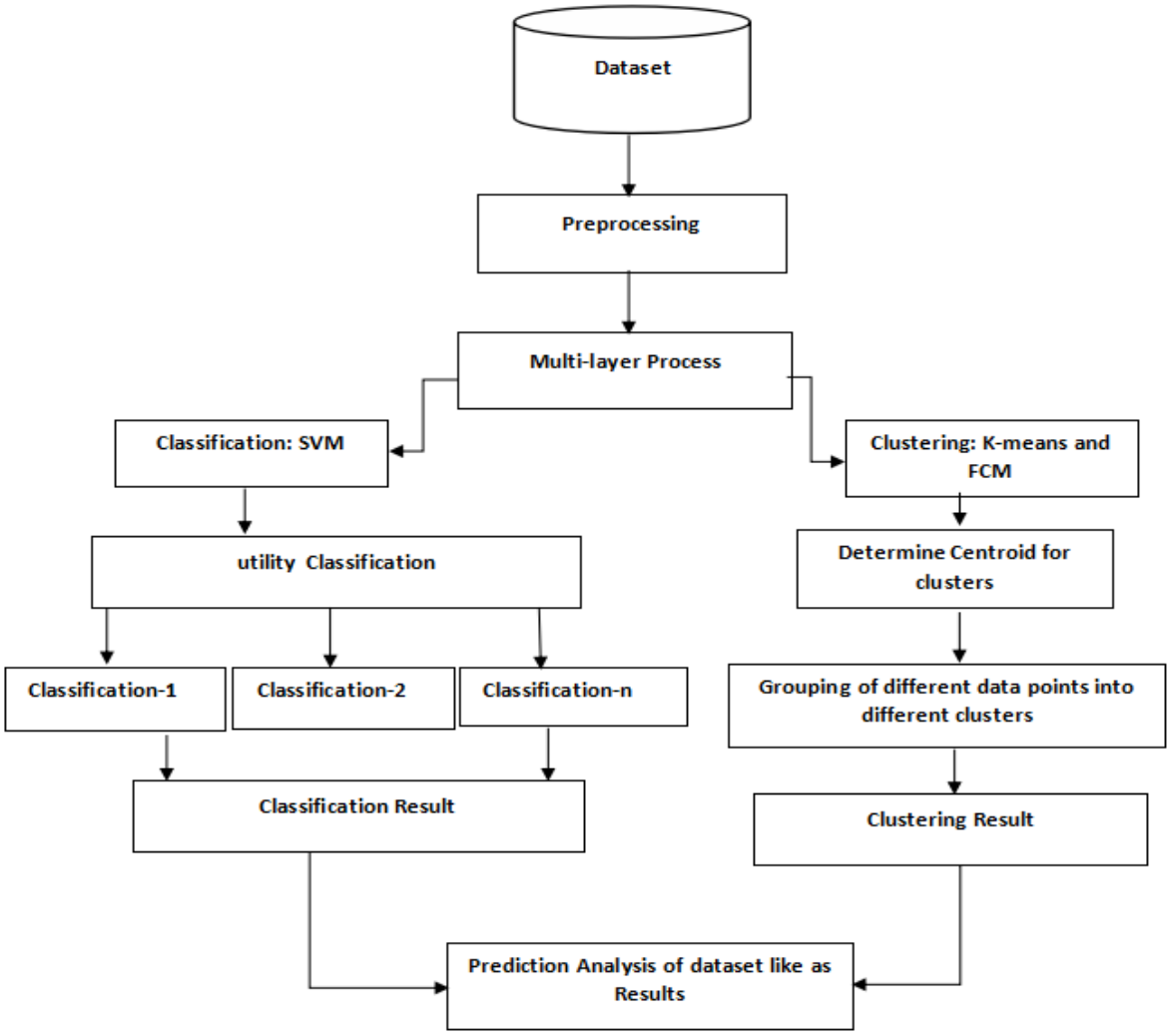

Fig.1. Workflow

5.RESULT AND DISCUSSIONS

A. CLUSTERING:

Cluster analysis is grouping the data objects based only on information found in the description of objects and their relationships. The goal of it is that the objects in one group are similar to each other(related) and different groups of objects are different(not related).The greater similarity within the group, the greater the difference between the groups, the better clustering. Clustering, as one of the methods of data mining, has been widely used in pattern recognition, trend analysis, similarity search and other areas. **[1].**

B.DATASET:

A dataset is a gathering of information. Most usually an informational collection compares to the substance of a solitary database table, or a solitary measurable information framework, where each segment of the table speaks to a specific variable, and each line relates to a given individual from the informational collection being referred to. Informational collections that are large to the point that customary information preparing applications are deficient to manage them are known as large information.

C.EFFICIENCY ANALYSIS:

Efficiency is frequently mistaken for adequacy. Effectiveness is a quantifiable idea, quantitatively controlled by the proportion of valuable yield to add up to enter. Viability is the more straightforward idea of having the capacity to accomplish a coveted outcome, which can be communicated quantitatively, however doesn't as a rule require more muddled arithmetic than expansion. Proficiency can regularly be communicated as a level of the outcome that could in a perfect world be normal, for instance if no vitality were lost because of grating or different causes, in which case 100% of fuel or other info would be utilized to deliver the coveted outcome.

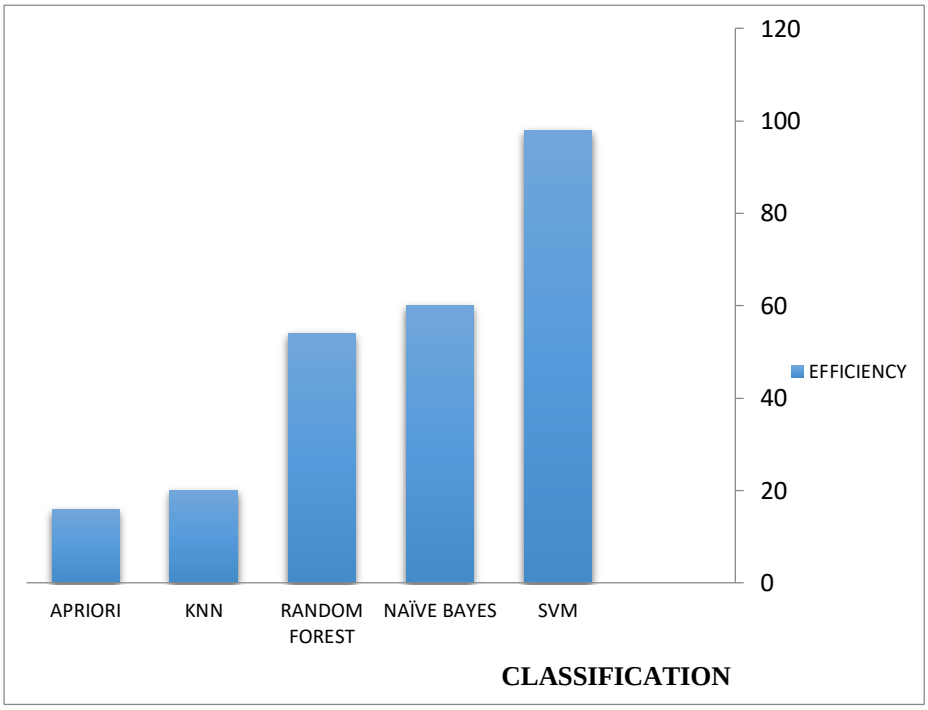

Fig.2. Efficiency Analysis of Classification

The Figure.2.discusses the efficiency of the classification process Since the effectiveness is a quantifiable idea, quantitatively controlled by the proportion of valuable yield to add up to enter, among various classification process more and more efficiency is shown , as it has the possibility to classify all various categories of data.

6.CONCLUSION:

In Data Mining, Association Rule Mining is one of the most important tasks. A large number of efficient algorithms are available for association rule mining, which considers mining of frequent itemsets. But an emerging topic in Data Mining is Utility Mining, which incorporates utility considerations during itemset mining. Utility Mining covers all aspects of economic utility in data mining and helps in detection of rare itemset having high utility. Rare High Utility itemset mining is very beneficial in several real-life applications. In this paper, we have presented a brief overview of various algorithms for high utility rare itemset mining. In the future scope, we will be presenting a comparative study of various algorithms for mining rare high utility itemset.

7. REFERENCES

1. R. Agrawal, T. Imielinski and A. Swami, 1993, "Mining association rules between sets of items in large databases", in Proceedings of the ACM SIGMOD International Conference on Management of data, pp 207-216.

2. R. Agrawal and R. Srikant, 1994, "Fast Algorithms for Mining Association Rules", in Proceedings of the 20th International Conference Very Large Databases, pp. 487-499.

3. Attila Gyenesei, "Mining Weighted Association Rules for Fuzzy Quantitative Items", Lecturenotes in Computer Science, Springer, Vol. 1910/2000, pages 187-219, TUCS Technical Report No.346, ISBN 952-12-659-4,ISSN 1239- 1891, May 2000.

4. R. Chan, Q. Yang, Y. D. Shen, "Mining High utility Itemsets", In Proc. of the 3rd IEEE Intel. Conf. on Data Mining (ICDM), 2003.

5. H. Yun, D. Ha, B. Hwang, and K. Ryu. "Mining association rules on significant rare data using relative support". Journal of Systems and Software, 67(3):181–191, 2003.

6. H.Yao, H. J. Hamilton, and C. J. Butz, "A Foundational Approach to Mining Itemset Utilities from Databases", Proceedings of the Third SIAM International Conference on Data Mining, Orlando, Florida, pp. 482-486, 2004.

7. G. Weiss. "Mining with rarity: a unifying framework",.SIGKDD Explor. Newsl., 6(1):7– 19, 2004.

8. Liu, Y., Liao, W., and A. Choudhary, A., "A Fast High Utility Itemsets Mining Algorithm", In Proceedings of the Utility- Based Data Mining Workshop, August 2005.

9. Lu, S., Hu, H. and Li, F. 2005. "Mining weighted association rules. Intelligent Data Analysis", 5(3):211–225.

10. V. S. Tseng, C.J. Chu, T. Liang, "Efficient Mining of Temporal High Utility Itemsets from Data streams", Proceedings of Second International Workshop on Utility-Based Data Mining, August 20, 2006

11. H. Yao, H. Hamilton and L. Geng, "A Unified Framework for Utilty-Based Measures for Mining Itemsets", In Proc. of the ACM Intel. Conf. on Utility-Based Data Mining Workshop (UBDM), pp. 28-37, 2006.

12. A. Erwin, R.P.Gopalan and N. R. Achuthan, 2007, "A Bottom-up Projection based Algorithm for mining high utility itemsets", in Proceedings of 2nd Workshop on integrating AI and Data Mining(AIDM 2007)", Australia, Conferences in Research and Practice in Information Technolofy(CRPIT),Vol. 84.

13. J. Hu, A. Mojsilovic, "High-utility pattern mining: A method for discovery of high-utility item sets", Pattern Recognition 40 (2007) 3317 – 3324.

14. L. Szathmary, A. Napoli, P. Valtchev, "Towards Rare Itemset Mining" Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, 2007, Volume 1, Pages: 305-312, ISBN ~ ISSN:1082-3409 , 0-7695- 3015-X

15. Kriegel, H-P et al. 2007. "Future Trends in Data Mining, Data Mining and Knowledge Discovery", 15:87–97.

16. M. Adda, L. Wu, Y. Feng, "Rare Itemset Mining", Sixth International conference on Machine Learning and Applications, 2007, pp 73- 80.

17. H.F. Li, H.Y. Huang, Y.Cheng Chen, and Y. Liu and S. Lee, "Fast and Memory Efficient Mining of High Utility Itemsets in Data Streams", 2008 Eighth IEEE International Conference on Data Mining.

18. M. Sulaiman Khan, M. Muyeba, Frans Coenen, 2008. "Fuzzy Weighted Association Rule Mining with Weighted Support and Confidence Framework", to appear in ALSIP (PAKDD),pp. 52-64.

19. S. Shankar, T.P.Purusothoman, S.Jayanthi and N.Babu, "A Fast Algorithm for Mining High Utility Itemsets", Proceedings of IEEE International Advance Computing Conference (IACC 2009), Patiala, India, pages : 1459 - 1464

20. Hu, J., Mojsilovic, A. "High-utility Pattern Mining: A Method for Discovery of Highutility Item" Sets, Pattern Recognition, Vol. 40, 3317-3324.

21. G.C.Lan, T.P.Hong and V.S. Tseng, "A Novel Algorithm for Mining Rare-Utility Itemsets in a Multi-Database Environment"

22. J. Pillai, O.P. Vyas, S. SoniM. Muyeba "A Conceptual Approach to Temporal Weighted Itemset Utility Mining", 2010 International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 28