Survey on Clustering of Massive Customer Transaction Data

Mrs. Sonali L. Mortale, Mrs. Manisha Darak
[1] Student M.E.Computer , [2]Professor ,Computer Department
[1]Computer Department
[1] Siddhant College of Engineering, Pune, India

***Abstract* — Today a clustering of customer transaction data is very important procedure and to analyze customer behaviors in retail and e-commerce companies. Product from companies is organized as product tree, in which the leaf nodes are goods to sell, and the internal nodes (except root node) could be multiple product categories. We propose the "personalized product tree", named purchase tree, to represent a customer's transaction records. Customer's transaction data set can be compressed into a set of purchase trees. We also propose a partitioned clustering algorithm, named PurTreeClust, for fast clustering of purchase trees. To cluster the purchase tree data, we first rank the purchase trees as candidate representative trees with a novel separate density, and then select the top k customers as the representatives of k customer groups. We also propose a gap statistic based method to evaluate the number of clusters. We use C 4.5 algorithm for making a decision tree, which can show different transaction of customer to make better purchase decision. Finally, the clustering results are obtained by assigning each customer to the nearest representative.**

Index Terms— **Customer segmentation, clustering transaction data, purchase tree, clustering trees.**

I. INTRODUCTION

Clustering is an important data mining technique that groups together similar data records. Recently, more attention has been put on clustering categorical data, where records are made up of non-numerical attributes[1]. Transactional data, like market basket data and web usage data, can be thought of a special type of categorical data having Boolean value, with applications in retail industry, e-commerce intelligence, etc. However, fast and effective clustering of transactional databases is extremely difficult because of the high dimensionality, sparsity, and huge volumes often characterizing these databases [2]. The segmentation of customers is very important for the enterprises of retail trade and electronic commerce, since that is usually the first step towards the analysis of the behaviors of the customers in these companies. The early work used general variables like customer demographics, and lifestyles, but such jobs are dubious because the general variables are difficult to collect and some of the variables collected may not be valid soon without fix. With the strategy to increase the behavioral data of the customers collected, researchers now focus on grouping customers based on the data of the transactions. The transaction data are the recorded information of the daily transactions of the customers, in that a transaction log contains a set of products (items) bought by a customer in a basket.

There exist three problems for clustering of customer transaction data: 1) how to represent a customer with the associated transaction records, 2) how to compute the distance between different customers, and 3) how to segment the customers into specific number of customer groups. Recently, Hsu et. el. proposed a customer clustering method of transaction data . In their method, the elements are organized in a hierarchical tree structure which is called in the concept hierarchy. Define a similarity measure based on the length of the route and the depth of the path in the hierarchy of the concept, and use the hierarchical clustering to segment customers. However, the distance is defined in the record of the individual transaction, so the method suffers from the huge amount of transaction records. On the other hand, the high computational complexity of the hierarchical clustering makes it difficult to use their methods of customer segmentation in real-world applications. They propose an algorithm of clustering PurTreeClust for the segmentation of customers from bulk data of customer transactions [1][3]. The remaining part of the paper includes: Section II contains literature review about the previous work and section III concludes the paper.

II. REVIEW OF LITERATURE

Xiaojun Chen, Yixiang Fang, Min Yang, Feiping Nie, Zhou Zhao and Joshua Zhexue Huang [1], the author propose a "personalized product tree", called tree-of-purchase, to represent the records of transactions of a customer. Therefore, the data set of customer transactions can be compressed into a set of trees of purchase. Propose an algorithm of partitioned clustering algorithm, named PurTreeClust for the rapid clustering of trees of purchase. A new distance metric is proposed to effectively calculate the distance between two buy trees. To cluster the data in the tree-of-purchase they first rank the purchase trees as candidate representative trees with a novel separate density, and then select the top k customers as the representatives of k customer groups. Finally, the clustering results are obtained by assigning each client to the nearest representative.

Yiling Yang Xudong Guan Jinyuan You [2], in this paper author studies about the problem of categorical data clustering, especially for transactional data characterized by high dimensionality and large volume. They start from the heuristic method of increasing the height-to-width ratio of the cluster histogram, they develop a novel algorithm – CLOPE, which is very fast and scalable, while being quite effective. They examine the algorithm on two real world datasets, and compare CLOPE with the state-of-art algorithms.

Kishana R. Kashwan and C. M. Velu [3], Clustering technology is an important step in data mining. The multivariate procedure is quite suitable for dividing the market, forecasting, planning and researching. This research, a comprehensive developed model of K-means Clustering technology and SPSS tools for developing real-time and online systems for specific supermarkets, has received direct input from the sales data record, and is now available in the market. Intelligent automatically updated segment statistics at the end of the day business.

Xiaojun Chen, Joshua Z. Huang, Qingyao Wu [4], A new algorithm for high-dimensional gene expression data Subspace Weighting Co-Clustering (SWCC) is proposed. A sub-space weight matrix was introduced to specific contributions to the identification of different clusters of samples of relics. They designed a new co-cluster objective function to recover co-clusters of gene expression data using subspace weight matrices. An iterative algorithm for solving the objective function is developed. In comparison with six advanced clustering algorithms on 10 sets of gene expression data sets, promising results of the proposed algorithm are demonstrated in the empirical study.

Alex Rodriguez, Alessandro Laio [6], the author propose an approach based on the idea that the cluster center is density than the neighborhood and the distance from the dense point is relatively large. The idea is based on a cluster method in which the number of clusters is intuitively generated, outliers are automatically detected and excluded from the analysis, and the clusters are recognized regardless of their shape and the dimension of the embedded space. It is a powerful algorithm for multiple test cases.

Table 1 Literature Survey

Sr. No.	Paper Title	Author	Method Proposed	Disadvantages
1	PurTreeClust: A Clustering Algorithm for Customer Segmentation from Massive Customer Transaction Data	Xiaojun Chen, Yixiang Fang, Min Yang, Feiping Nie, Zhou Zhao and Joshua Zhexue Huang,	Propose a “personalized product tree”, named purchase tree, to represent a customer’s transaction records.	Not perfect decision tree.
2	Clope: a fast and effective clustering algorithm for transactional data	Y. Yang, X. Guan, and J. You	This paper studies the problem of categorical data clustering.	Performance of this method is not good.
3	Customer Segmentation Using Clustering and Data Mining Techniques	Kishana R. Kashwan and C. M. Velu	This research paper is a comprehensive report of k-means clustering technique and SPSS Tool to develop a real time and online system for a particular super market to predict sales in various annual seasonal cycles.	High computational cost.
4	Subspace weighting co-clustering of gene expression data	X. Chen, J. Z. Huang, Q. Wu, and M. Yang	They propose a novel algorithm, called Subspace Weighting Co-Clustering (SWCC), for high dimensional gene expression data.	Effectiveness of system is not good.
5	TW-k-means: Automated Twolevel Variable Weighting Clustering Algorithm for Multi-view Data	X. Chen, X. Xu, Y. Ye, and J. Z. Huang	This paper proposes TW-k-means, an automated two-level variable weighting clustering algorithm for multiview data.	Poor accuracy.
6	A. L. Alex Rodriguez	Clustering by fast search and find of density peaks	They propose an approach based on the idea that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities.	The density estimated for a finite sample size is far from uniform and is instead characterized by several maxima.

III. PROPOSED SYSTEM

We build a product tree, where the leaf nodes usually denote the provided products and the internal nodes are multiple product categories. A product tree often consists of several levels and thousands of nodes, and the number of nodes increases rapidly from the top level to the bottom level. In transaction data, each product (item) bought by a customer corresponds to a leaf node in the product tree. To facilitate the analysis and visualization of customer's behavior, we build a ”personalized product tree” for each customer, called purchase tree. The purchase tree can be built by aggregating all products in a customer's transactions, and pruning the product tree by only retaining the corresponding leaf nodes and all paths from root to leaf node. The tree edit distance will produce high distance value between any two purchase trees because customers do not buy similar products. Therefore, it is difficult to recover cluster structure with the tree edit distance. To solve this problem, we have to effectively utilize the semantic in the product tree. We define a new PurTree distance metric to compare customers from the entire levels of the product tree. The most important property of the new distance is that it is a metric, thus many advanced techniques can be used for the purchase tree data with the new distance. We propose a fast density estimation method for purchase tree data, and a separate density to rank the purchase trees as candidate representative trees which are both dense and far from each other.

IV. CONCLUSION

This paper discussed about the clustering of customer transaction data. Clustering is an important data mining technique that groups together similar data records. They have presented PurTreeClust for massive customer transaction data. In this “purchase tree” is built for each customer from the customer transaction data. A new distance metric is defined to effectively compute the distance from two purchase trees. Also present gap statistic based method to evaluate the number of clusters.

REFERENCES

[1] Xiaojun Chen, Yixiang Fang, Min Yang, Feiping Nie, Zhou Zhao and Joshua Zhexue Huang, “PurTreeClust: A Clustering Algorithm for Customer Segmentation from Massive Customer Transaction Data”, 2017.
[2]Y. Yang, X. Guan, and J. You, “Clope: a fast and effective clustering algorithm for transactional data,” 2002.
[3] Kishana R. Kashwan and C. M. Velu,“Customer Segmentation Using Clustering and Data Mining Techniques”, 2013.
[4] X. Chen, J. Z. Huang, Q. Wu, and M. Yang, ‘Subspace weighting co-clustering of gene expression data.,”2017.
[5] X. Chen, X. Xu, Y. Ye, and J. Z. Huang, “TW-k-means: Automated Twolevel Variable Weighting Clustering Algorithm for Multi-view Data,” 2013.
[6] A. L. Alex Rodriguez, “Clustering by fast search and find of density peaks,” 2014.
[7] T. Xiong, S. Wang, A. Mayers, and E. Monga, “Dhcc: Divisive hierarchical clustering of categorical data,” 2012.
[8] V. L. Migu´eis, A. S. Camanho, and J. F. e Cunha, “Customer data mining for lifestyle segmentation,.” 2012.
[9] M. Pawlik and N. Augsten, “Rted: a robust algorithm for the tree edit distance,” 2011.
[10] R. Kuo, L. Ho, and C. M. Hu, “Integration of self-organizing feature map and k-means algorithm for market segmentation,.” 2002.