International Journal of Computer Applications Technology and Research
Volume 4– Issue 2, 126 - 128, 2015, ISSN:- 2319–8656

A Survey on the Clustering Algorithms in Sales Data Mining

Mathew Ngwae Maingi
School Of Computing And Information Technology
Jomo Kenyatta University Of Agriculture And Technology
Nairobi, Kenya

Abstract: This paper discusses different clustering techniques that can be used in sales databases. The advancement of digital data collection and build up of data in data banks as a result of modernization in sales disciplines has brought in great challenges of data processing for better and meaningful results due to mass data deposits. Clustering techniques therefore are quite necessary so that the senior management in sales department can have access to processed data as they engage themselves in decision making processes. In this paper, I focus on the retail sales data mining, classification and clustering techniques. In this study I analyze the attributes for the prediction of buyer's behavior and purchase performance by use of various classification methods like decision trees, C4.5 algorithm and ID3 algorithm.

Keywords: clustering; databases; banks; discipline; management; ID3; algorithm; C4.5.

1. INTRODUCTION

In sales database systems, there exist many varieties of functions for handling many processes such as supply chain management, marketing strategies, market analysis performance in identifying new product issues, diagnosis of manufacturing problems causes and profiling existing customers with more accurate and tangible values. This huge collection of data values is either related or not related at all and thus definitely needs to be clustered otherwise much of the data deposited will not be useful to users. From data mining perspective, clustering method plays a very crucial role in knowledge discovery in such activities as cross marketing: increase the sales in season wise by updating the inventory, discount offers and store layout based on the knowledge discovered in the data.

2. WHAT IS A CLUSTERING ALGORITHM IN SALES DATA?

In today's world, an enterprise that processes over 15 million point-of-sale transactions a day in its database would most likely find that data of less use without analyzing it using data mining software. If that data from the point-of-sale system was properly analyzed using data mining techniques, this will enhance accurate: determination of sales trends, development of marketing campaigns, and prediction of customer dependability. Clustering algorithms therefore seek to group given data sets into groups based on identified features so that the data points within a group are more similar to each other than the points in different groups.

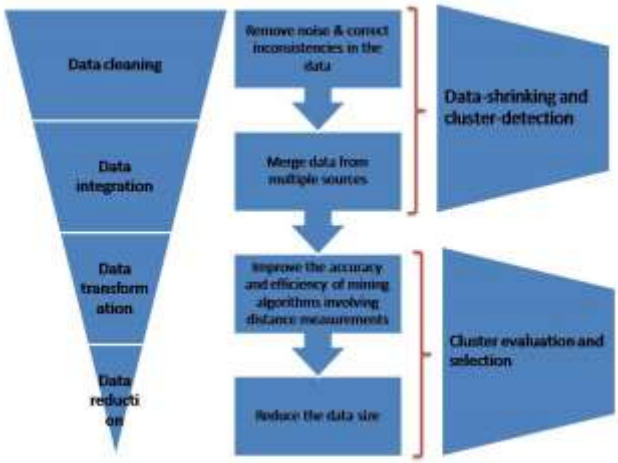

Figure 1 Data processing techniques in cluster formation.

In the figure 1 above, it is shown that there is need to clean data by removing the available outliers inorder to prepare it for use. Data integration entails merging sata from multiple sources so as to centralize it. Data transformation deals with accuracy and efficiency improvement of data mining algorithm hence fine tuning the resultant data. As a result, the sales data set can be extracted from sales transactions since these records are directly linked to the entire sales data. With such a technique, data values can be given as input data where else on the other hand in output it gives recommendations to management functions and extract new knowledge to the managers by using various data mining techniques like clustering, classification or pattern matching.

2.1 Goals and objectives clustering on sales data

Clustering seeks to group given data sets into groups based on identified features so that the data points within a group are more similar to each other than the points in different groups. The following are some of the advantages:

1. Prediction of customer purchasing behavior by grouping similar data points together into clusters hence generating knowledge based models.

2. Upgrading and advancing scientific knowledge discover through tangible analysis.

3. STAGES OF SALES DATA CLUSTERING

Sales data clustering is concerned with group given data sets into groups based on identified features. The process generally consists of three phases:

1. Feature selection and clustering algorithm formation. This phase deals with proper selection of the features on which the clusters are to be formed such as distance definition, clusters to be formed e.t.c.
2. Validation of the results that deals with the assessment of the quality and the reliability of the cluster sets. Clustering algorithms yield results that are not predictable despite the method used and therefore the final partition of data may need reevaluation.
3. Interpretation of the results. In the process of deriving the conclusion of any output, experts from the application areas integrates the clustering results with other experimental output and do the necessary analysis.

3.1 Clustering methods

There are various methods of clustering sales data where each uses a unique induction principle. All kinds of methods fall under one of the following subcategories:

1. **Hierarchical:** This category of clustering seeks to construct a hierachy of clusters using such strategies as: agglomerative and divisive. Agglomerative approach uses "bottom up" strategy where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive strategy is a "top down" approach where all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. The major overall limitations of the hierarchical methods are: Inability to scale well—The time complexity of hierarchical algorithms is at least O(m2). Clustering a large number of objects using a hierarchical algorithm is also characterized by huge I/O costs. Hierarchical methods can never undo what was done previously. Hence lack of back-tracking capability. This approach above make the derived hierarchical tree more robust, however, it doesn't indicate how to cut the dendrogram to obtain meaningful clusters, either.
2. **Partitioning:** These methods relocate instances by moving them from one cluster to another starting from an initial partitioning. To achieve overall optimality in this type of clustering, an exhaustive enumeration process of all possible partitions is required. Because this is not feasible, certain greedy heuristics are used in the form of iterative optimization. The following subsections present various types of partitioning methods.
3. **Grid-based Algorithms:** In this method, partitions are created in the space to form finite number of cells that form a grid like structure where all clustering operations are performed. This method however has one advantage of being very fast in processing time regardless of the data size though dependent on the number of segments in each dimension in the quantized space. In the case of spatial data mining, the approach has been improved to reduce cost. Study shows that this approach performs very efficiently in the case where the data sets are very large.
4. **Density-based algorithms:** This method assumes that the points belonging to a specific cluster are drawn from a specific probability distribution. More so, the method describes the distribution of a data set by the density of the data objects hence, the entire process involves the search of the dense areas in the object space. The main aim of this approach is to identify clusters and their distribution parameters and is designed to discover clusters for arbitrary shape that are not convex. This approach is characterized by presence of noise within the data sets and this might interfere with the overall process of data mining. The approach also capitalizes on the ability to discover clusters with arbitrary shape and has good efficiency on large data sets, however, the approach is very sensitive to the input parameters hence it is prone to generating very different clustering results if subjected to slightly different parameter settings.

4. LITERATURE SURVEY

4.1 Predicting customer purchase in an online retail business, a Data Mining approach

In this research, Aniruddha Mazumdar, May 2010, studied, implemented and analyzed some data mining tools as well as techniques and later analyzed the sampled data for interpretation. In this study data mining algorithms were used based on the clustering algorithms in conjunction with an 'Apriori' based Association rule mining algorithm. The research discusses different approaches that were used in interpreting the results. The results clearly prove that:

Using the VQ approach one can easily segment groups of buyers based on the RFM values or all of them all together. However this process needs the initials vectors as its input for it to form clusters.

1. When predicting, different segmentations from the clusters formed can be used and the most populous ones are specifically focused on.
2. Basing on the association rules alongside sufficient coverage, the product which the customer wishes to buy is predictable along with the purchase of particular products.

4.1 Upgrading and advancing scientific knowledge discovery through tangible market analysis.

In this paper by Kavitha .N et al 2013, it is argued that despite the fact that there have been several techniques, like pattern discovery, association rule mining etc, these methods generates a large volume of frequent patterns and rules which are not useful for finding the essential patterns among them,

from the database. Therefore the discovery of hidden information present in databases and can be regarded as a step in overall process of Knowledge Discovery in databases (KDD) (Parashar, 2011)

A research paper by D. Bhanu, Dr. S. Pavai Madeshwari , 2009 proposes an architecture to be used to discover a customer-based rules if a retailer want to open his outlet at an entirely new location. However, if these rules were to be obtained, then a fuzzy clustering method would be used for customer and product domains and be bridged. Association rule mining and Fuzzy clustering get incorporated to analyze the similarity between customer groups and their preferences for products. Once a complete set of rules is generated, this is put in an independent knowledgebase from which any stated or needed customer needs can be categorized with the correspondence to customer groupings hence deducing the cluster to which a customer may belong.

5. CONCLUSION

This research paper describes the sales knowledge discovery, objectives and goals of sales knowledge discovery as well as the stages or phases of sales knowledge discovery together with the existing techniques of classification. It is clear that various techniques of classification can be implemented on the data set however it is worth noting that, the technique of classification to be applied on the data to improve customer analysis in purchasing is very important.
So this paper will provide a beneficial overview of existing solutions for classification and clustering methods with their advantages, limitations as well as the strengths of data clustering in decision making in an enterprise.

6. ACKNOWLEDGMENTS

I would like to express my sincere gratitude to almighty God, my supervisor Dr. Wilson Cheruiyot and his colleagues for helping me to undertake this research from start to completion. God bless you all.

7. REFERENCES

[1] Kaviha N. and Karthikeyan S. 2013. Customer Buying Behavior Analysis: A Clustered Closed Frequent Itemsets for Transactional Databases.
[2] Aniruddha Mazumdar. 2010. Predicting customer purchase in an online retail business-a data mining approach.
[3] D. Bhanu, Dr. S. Pavai Madeshwari 2009. Retail market analysis in targeting sales based on consumer behavior using Fuzzy Clustering – A rule based model.
[4] Wei Li, 2008. Modified K-means clustering algorithm IEEE computer society Congress on Image and Signal Processing, (May 2008), 618-621
[5] James Shields 2008. Getting to Know Your Customers by Clustering on Product Purchase Patterns