International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 07, July-2018

Improving Sales Analysis in Retail Sale using Data Mining Algorithm with Divide and Conquer Method

Myint Myint Yee
Department of Information Technology Supporting and Maintenance
University of Computer Studies, Yangon
Yangon, Myanmar

***Abstract*— Data is recognized as a basic form of information that needs collection, management, mining and interpretation to create knowledge. The advantages of the collection of digitalized data and build data banks has brought in great challenges of data processing for better and meaningful results according to mass data deposits. Clustering techniques therefore are necessary in sales data analysis which can have access to process data in decision making processes. The purpose of this article is to present some basic ideas on sales analysis that support the process of determining the minimum stock and profit margin by grouping items into categories "Fast moving" and "Slow moving" and "Dead Stock" of the sale using clustering algorithm. In order to cluster the items, this proposed system will use k-mean clustering algorithm, divide and conquer method.**

Keywords*— *K-Means, Data Mining, Data Cube, Divide and Conquer

I. INTRODUCTION

Data Mining is the process of extracting useful information and patterns from enormous data. Data Mining includes collection, extraction, analysis and statistics of data. It is also known as Knowledge discovery process, Knowledge Mining from Data or data/ pattern analysis. Data Mining is a logical process of finding useful information to find out useful data. Once the information and patterns are found it can be used to make decisions for developing the business. Data mining tools can give answers to the various questions related to business which was too difficult to resolve. They also forecast the future trends which lets the business people to make proactive decisions. This paper test on the dataset that contains several categories such as Beer, Liquor & Wine, Beverage, Books and Music, Cosmetic and Toiletries, Delicatessen, Electric, Frozen Food & Frozen Meat, Fruit & Vegetable, Grocery, Kitchen Ware, Linens & Furniture and Stationery etc. If that huge collection of data values was properly analyzed using data mining techniques, this will enhance accurate: determining of sale trends, development of marketing campaigns, and prediction of customer dependability. Clustering is one of the importance functionality of the data mining and the process of grouping the data into classes or cluster, so that objects within a cluster have high similarity in comparison to one another and very similar to object in other clusters. Dissimilarity is due to the attributes values that describe the objects. In this paper, we use the large sales data that includes the sales records for miscellaneous items in diverse locations. We cluster the categories "Fast moving" and "Slow moving" and "Dead Stock" of the sale which can be used later for decision making and improving the activities of the business by using data mining algorithm. We propose an efficient algorithm that is based on divided and conquers techniques for clustering the large sale datasets. In this paper, we need to collecte the data for analysis which consist of data of sales per quarter, for the years 2015 to 2017. We are, however, interested in the sales by product (no. of transition, total revenue, profit margin). Thus the data can be summed up so that the resulting data summarize the total sale by product ready for data mining.

II. RELATED WORKS

Kusrin Kusrini studied to support the process of determining the minimum stock and profit margin by building a model that can group items into categories "fast moving" and "slow moving" using only k-means clustering algorithm[1].

In [2], the researchers proposed a method that uses divide and conquer technique with equivalency and compatible relation concepts to improve the performance of the K-Means clustering method for using in high dimensional datasets. Experiment results demonstrated appropriate accuracy and speed up.

M. N. Maingi analyzed 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 [3].

A researcher S.T. Deokar discusses the implementation of K-Means clustering algorithm for clustering unstructured text documents that implemented, beginning with the representation of unstructured text and reaching the resulting set of clusters. In this work, Residual sum of square, Tf-IDF is also used [5].

Some researchers in [6] tested on dataset of retail sales using WEKA interface and computer the correctly cluster building instances in proportion with incorrectly formed cluster. They compared four algorithms of incorrectness percentages. They showed the result that simple K-Means algorithm formed higher incorrectly cluster instances than other 3 algorithms in more clusters.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 07, July-2018

III. BACKGROUND THEORY

A. *Sales Analysis*

Sales analysis examines sales reports to see what goods and services have and have not sold well. The analysis is used to determine how to stock inventory, how to measure the effectiveness of a sales force, how to set manufacturing capacity and to see how the company is performing against its goals.

A sales analysis report comprises of a company's sales volume trends over time. The basic purpose of a sales analysis report is to determine whether sales are increasing or going down. During a fiscal year, sales managers, at any point of time, may analyze the trends in the sales analysis report to determine future strategies for paving forward the best course of action. Sales analysis reports are often used by sales managers to identify market opportunities and areas where the potential of increasing sales volume lies.

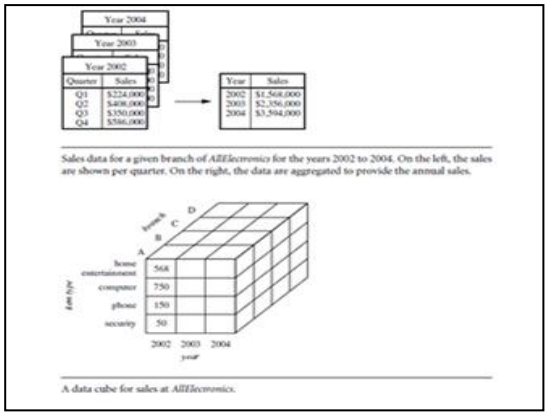
B. *Data Cube Aggregation*

Complex data analysis and mining on huge amounts of data can take a long time, making such analysis impractical or infeasible.

Data reduction techniques can be applied to obtain a compressed representation of the data set that is much smaller in volume, yet maintains the integrity of the original data.

Strategies for data reduction include:

1. Data cube aggregation—aggregation operations are applied to the data in the construction of a data cube.
2. Attribute subset selection—irrelevant, weakly relevant or redundant characteristics or dimensions may be detected and removed.
3. Dimensionality reduction,—encoding mechanisms are used to reduce the dataset size.
4. Numerosity reduction—data is replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data) or nonparametric methods such as clustering, sampling, and the use of histograms.
5. Discretization and concept hierarchy generation—raw data values for attributes are replaced by series or higher conceptual levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies. Discretization and concept hierarchy generation are powerful tools for data mining, in that they allow the mining of data at multiple levels of abstraction.
6. The computational time spent on data reduction should not outweigh or erase the time saved by mining on a reduced data set size.



Source: http://www.faadooengineers.com/notes/images/1/56/986ad00660e6b532f524c778104a95bc1.png
Figure 1: Sample Data Cube

This aggregation is illustrated in the figure above. The resulting data set is smaller in volume, without loss of information necessary for the analysis task.

Data cubes store multidimensional aggregated information. Each cell holds an aggregate data value, corresponding to the data point in multidimensional space. Concept hierarchies may exist for each attribute, allowing the analysis of data at multiple levels of abstraction. For example, a hierarchy for branch could allow branches to be grouped into regions, based on their location. Data cubes provide fast access to precomputed, summarized data, benefiting on-line analytical processing as well as data mining.

The cube created at the lowest level of consideration is referred to as the base cuboid. The base cuboid should resemble an individual entity of interest, such as sales or customer. In other words, the lowest level should be used for the analysis. A cube at the highest level of abstraction is the apex cuboid.

Data cubes created for varying levels of abstraction are often referred to as cuboids, so that a data cube may instead refer to a lattice of cuboids. Each higher level of abstraction further reduces the resulting data size. When replying to data mining requests, the smallest available cuboid relevant to the given task should be used.

C. *Cube Mining Techiques*

One of the most important tasks in Data Mining is to select the correct data mining technique. Data mining technique has to be chosen based on the type of business and the type of problem of business faces. A generalized approach has to be used to improve the accuracy and cost effectiveness of using data mining techniques. There are basically seven main Data Mining techniques. There are also a lot of other Data Mining techniques but these seven are considered more frequently used by business people.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 07, July-2018

1. Tracking patterns. One of the most basic techniques in data mining is learning to recognize patterns in data sets. This is usually a recognition of some aberration in data happening at regular intervals, or an ebb and flow of a certain variable over time. For example, we might see that sales of a certain product seem to spike just before the holidays.

2. Classification. Classification is a more complex data mining technique that forces to collect various attributes together into discernable categories, which we can then use to draw further conclusions, or serve some function. For example, if we're evaluating data on individual customers' financial backgrounds and purchase histories, we might be able to classify them as "low," "medium," or "high" credit risks. We could then use these classifications to learn even more about those customers.

3. Association. Association is related to tracking patterns, but is more specific to dependently linked variables. In this case, we'll look for specific events or attributes that are highly correlated with another event or attribute; for example, we might notice that when our customers buy a specific item, they also often buy a second, related item. This is usually what's used to populate "people also bought" sections of online stores.

4. Outlier detection. In many cases, simply recognizing the overarching pattern can't give us a clear understanding of our data set. We also need to be able to identify anomalies, or outliers in data. For example, if our purchasers are almost exclusively male, but during one strange week in July, there's a huge spike in female purchasers, we'll want to investigate the spike and see what drove it, so we can either replicate it or better understand our audience in the process.

5. Clustering. Clustering is very similar to classification, but involves grouping chunks of data together based on their similarities. For example, we might choose to cluster different demographics of our audience into different packets based on how much disposable income they have, or how often they tend to shop at our store.

6. Regression. Regression, used primarily as a form of planning and modeling, is used to identify the likelihood of a certain variable, given the presence of other variables. For example, we could use it to project a certain price, based on other factors like availability, consumer demand, and competition. More specifically, regression's main focus is to help uncover the exact relationship between two (or more) variables in a given data set.

7. Prediction. Prediction is one of the most valuable data mining techniques, since it's used to project the types of data we'll see in the future. In many cases, just recognizing and understanding historical trends is enough to chart a somewhat accurate prediction of what will happen in the future. For example, we might review consumers' credit histories and past purchases to predict whether they'll be a credit risk in the future.

D. K-Means Clustering Algorithm

K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing anobjective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}\left(\|x_i - v_j\|\right)^2 \qquad (1)$$

where,

'$\|x_i - v_j\|$' is the Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers

Algorithmic steps for k-means clustering:

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center using:

$$v_i = (1/c_i)\sum_{j=1}^{c_i} x_i \qquad (2)$$

where,
'c_i' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 07, July-2018

E. K means Clustering Algorithm based on Divide and Conquer Technique

Since in each iteration K-Means algorithm computes the distance between data points and all centroids, this is computationally very expensive for large dataset, therefore we are using divide and conquer technique to reduce the number of computations which results in less execution time.

Input: Dataset D and number cluster, k.

Output: A set of k clusters

Divide Phase:

Step 1: Partition the dataset $\mathbf{D}_{rxd}$ into k parts as an average. At each iteration, $1<k<N/(4C)$, where C is the maximal number of clusters prospected and N is the length of the dataset.

Step 2: Apply K-Means algorithm on each of the partition of the dataset to get the initial clusters of each partition.

Step 3: Calculate the means of each clusters in all partitions separately.

Step 4: Calculate the average of all means within a partition.

Step 5: Repeat step 4 for all partitions.

Merge Phase:

Step 6: Now taking the average means from each partition as the final centroids for the final clustering, calculate the square of Euclidean distance of each data point in the dataset to the above average means.

Step 7: Finally based on the minimum distance criterion, assign each data points to the cluster to which it has minimum distance.

IV. PROPOSED SYSTEM

A. Proposed System Overview

Our proposed system is divided into 2 distinct phases describe as follows:

- The data is collected from the database and pre-processing is done on data. After the pre-processing is done aggregate the data by product to generate Data Cube.
- In second phase, we apply the hybrid clustering algorithm: K-Means with Divide and Conquer. Clusters are formed on bases of dead stock, slow moving and fast moving stock.

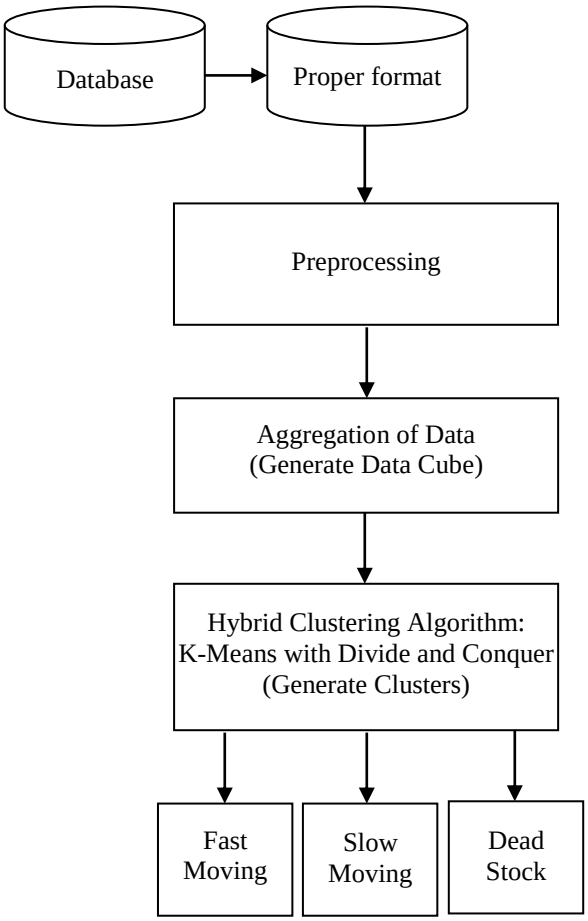


Figure 2: Overview of Proposed System

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 07, July-2018

B. Hybrid Algorithm: K- Means with Divide and Conquer

Abbreviations and Acronyms

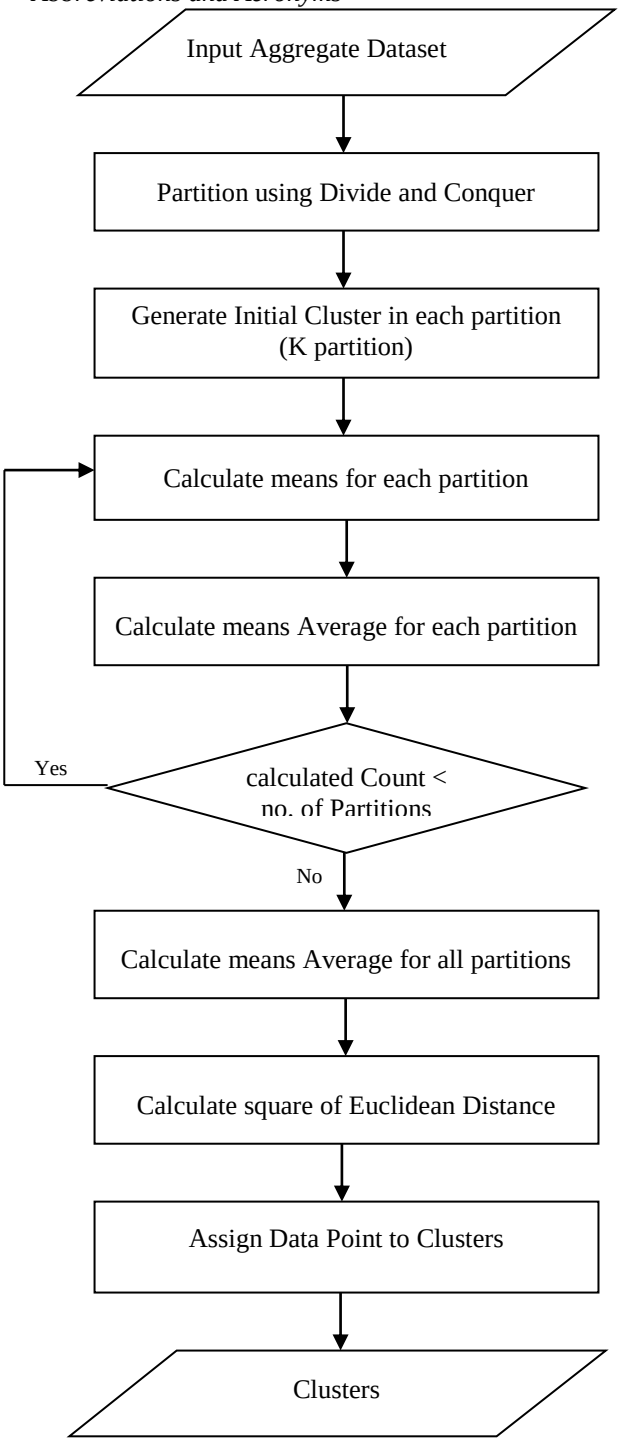


Figure 3: Flow of Hybrid Algorithm

V. EXPERMENTAL RESULT

Dataset are collected from City Mart Supermarket-Myanmar in 2015 to 2017 containing huge amount of sales data from different region. The instances in clusters are as shown in table 1.

Table 1: shows the instances of each cluster

Instances in Cluster1 (Fast moving)	Instances in Cluster2 (slow moving)	Instances in Cluster3 (Dead Stock)
1575	137	62

VI. EVALUATION

There are numerous evaluation measures to validate the cluster quality. To evaluate the clustering results of proposed system, precision, recall, and F-measure has been used.

For cluster *j* and cluster *i*:

$$\text{Recall } (i,j) = n_{ij}/n_i \quad (3)$$

$$\text{Precision } (i,j) = n_{ij}/n_j \quad (4)$$

where n_{ij} is the number of members of class *i* in cluster *j*, n_j is the number of members of cluster *j* and n_i is the number of members of class *i*. F-measure is computed using precision and recall as below:

$$F(i,j) = \frac{2 * recall\,(i,j) * precision\,(i,j)}{(precision\,(i,j) + recall\,(i,j)} \quad (5)$$

Table 2: shows the F-values for Hybrid Algorithm (K-Means + Divide and Conquer)

No. of Clusters(K)	F- measure
	Hybrid Algorithm (K-Means + Divide and Conquer)
K=3	0.66

VII. CONCLUSION

K-means algorithm with divide and conquer can be clustered in item grouping process into categories of fast moving and slow moving. By using the sales data in City Mart Supermarket Myanmar year 2015 and 2017, it is shown that the best cluster is occurred in clustering process with yearly data and attributes' values.

VIII. REFERENCES

[1] K. Kusrini, "Grouping of Retail Items by Using K-Means Clustering", Proceedings of The Third Information Systems International Conference, pp 495-502, 2015.

[2] M.Khalilian, F. Z. Boroujeni, N. Mustapha and Md. N. Sulaiman, "K-Means Divide and Conquer Clustering", Proceedings of International Conference on Computer and Automation Engineering, ICCAE/IEEE, 2009.

[3] M.Khalilian, N.Mustapha, Md. N. Sulaiman and Md. Ali Mamat, "A Novel K-Means Based Clustering Algorithm for High Dimentional Data Sets", Proceedings of the International MultiConference of Engineers and Computer Scientists 2010 vol 1, IMCES'10, 2010.

[4] M. N. Maingi,"A Survey on the Clustering Algorithms in Sales Data Mining",IJCATR. Vol 4 – Issue 2, 126-128, 2015.

[5] S.T.Deokar, "Text Documents Clustering using K Means Algorithm", Proceedings of International Journal of Technology and Engineering Science, IJTES. vol 1, 282-286, 2013.

[6] V.Shrivastrava and P.N. Arya, "A Study of Various Clustering Algorithms on Retail Sales Data", Proceedings of International Journal of Computing, Communications and Networking, IJCCN'12, 2012.