Implementation of Data Mining With Clustering of Big data for Shopping mall's data using SOM and K-means Algorithm

Fatema Jamnagarwala, Dr.P.A.Tijare

Student, Associate Professor Computer Engineering, Sipna Coet , Amravati,India

Abstract : *Study of customer behaviour in online shopping usually deals with identification of customers and their buying behaviour patterns. The aim of such studies is to make certain who buys where, what, when and how. The results of these studies are useful in the solution of marketing problems. Various studies on customer purchasing behaviours have been presented and used in real problems. For analysis of customer behaviours data mining techniques are consider more effective. The target of this paper is to analyze behaviour of such people who are visiting the online shopping sites and spending their time there, surfing for different stuff. It would also be taken into account that how many people are there and how many of them are actually shopping. In this paper, different queries are applied to mine the database of a specified site which results in analysis of customer behaviour towards online shopping.*

Keywords *— data mining; clustering, association rule; frequent item set;*

I. INTRODUCTION

Clustering is a technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. Data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields like e- commerce.

We can simply define data mining as a process that involves searching, collecting, filtering and analysing the data. It is important to understand that this is not the standard or accepted definition. But the above definition caters to the whole process.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

A. Big Data

Big Data itself termed as large datasets that their sizes are beyond the ability of capturing, managing, and processing by most software tools and people [1]. For example, Search engines, social networking, online Advertising, ecommerce, as well as education, healthcare, and medicine, etc. As the datasets are so long it have to face the challenges including capture, storage, [2] search, sharing, transfer, analysis and visualization. The trend are becoming to larger as the data sets are useful for the additional information deliverables from analysis of a single large set of related data, as compared to separate smaller sets with identical total quantity of knowledge, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases in medical sector, combat crime and determine real-time roadway traffic conditions or other real-time applications.

The advent of knowledge technology in varied fields of human life has crystal rectifier to the massive volumes of knowledge storage in varied formats like records, documents, images, sound recordings, videos, scientific data, and many new data formats. For better decision making, the data collected from different applications require proper mechanism of extracting useful knowledge/information from large data repositories [3,4].

B. Objective

The aim of data mining is to discover structure inside unstructured data, extract meaning from noisy data, discover patterns in apparently random data, and use all this information to better understand trends, patterns, correlations, and ultimately predict customer behaviour, market and competition trends, so that the company uses its own data more meaningfully to better position itself on the new waves. Here we are using customer shopping transaction which is in unstructured data but we have to mine data to retrieve some useful information from it.

II. LITERATURE SURVEY

1) Michael Steinbach from University of New York presents in his paper the results of an experimental study of some common document clustering techniques. In particular, comparison of the two main approaches to document clustering, agglomerative hierarchical clustering and K-means is done. Hierarchical clustering is often portrayed as the better quality clustering approach, but is limited because of its quadratic time complexity. In contrast, K-means and its variants have a time complexity which is linear in the number of documents, but are thought to produce inferior clusters.[4]

2) Charu C.Aggarwal and Chengxiang Zhai, in their paper provided a detailed survey of the problems of text clustering. Clustering is a widely studied data mining problem in the text domains. The problem finds numerous applications in customer segmentation, classification, collaborative filtering, visualization, document organization and indexing. [5]

3) Uses a distance metric for clustering high dimensional data based on the hitting time of two Minimal Spanning Trees (MST) grown sequentially from a pair of points by Prim's algorithm[6].

4) Yieng Chen and Bing Qin in their papers compared SOM and K means algorithm. K means is easy to realize and it usually has low computation cost, so it has become a well-known text clustering method. The shortcoming of K means is that the value of k must be determined before and initial documents points seeds need to be selected randomly. If the neuron number is less than the class number, it will not be sufficient to separate all the classes, the documents from some closely related class may be merged into one class. If the neuron number is more than the class number, the clustering results may be too fine. And the clustering efficiency and the clustering quality may also be adversely affected.[7]

III. IMPLEMENTATION AND WORKING METHODOLOGY

In our work we used an online shopping mall’s website’s database as a sample database. The site is about multiple items like Cosmetics, Utensils, Grocery, Vegetables, Fruits, Juices and Other. The site contains thousands of records that can give a little insight about customer purchasing habits in shopping malls. The reason for choosing this database was that, it has so many numbers of customers visiting every day that provides a huge dataset to analyse customer behaviour as Fig. 1 presents a methodology of our work[2,8].

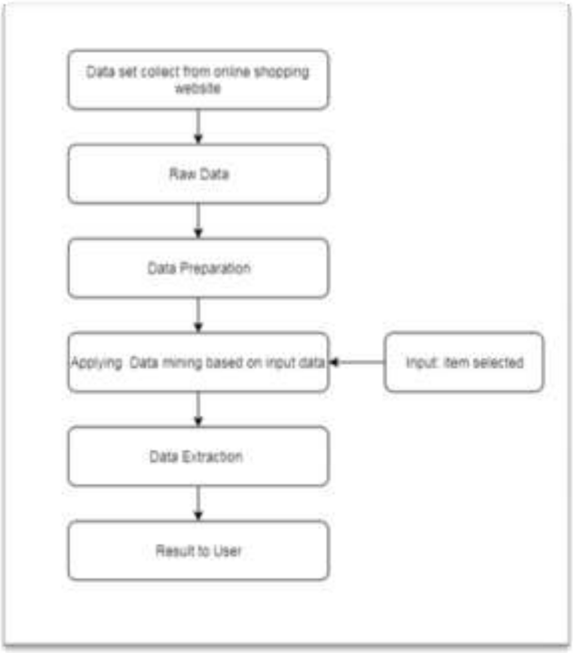

Data set collect from online shopping website
Raw Data
Data Preparation
Applying Data mining based on input data
Input item selected
Data Extraction
Result to User

Fig. 1. Implementation

To extract some useful pattern, after gathering such information, we select the reasonable strategy from several different choices. For analyzing customer shopping behavior there exist many old techniques such as conducting surveys through questionnaires and polling. Even various algorithms have been applied on online data for this kind of information. We here by apply association rule for checking the association between products which are bought by the customers. The basic idea here is to provide a mechanism of creating a market strategy out of some information retrieved from online purchasing.

Here when user selects any item as per their requirements based on given option from item data set. After selecting particular item our data mining algorithm will work based on selected input and generates and desired output which is best suggestion extracted from users transaction which are captured in past. Best suggestion can be generated by extracting best matching of item with selected input and only that items are given for the suggestions by sorting them with best match to least match clusters[9].

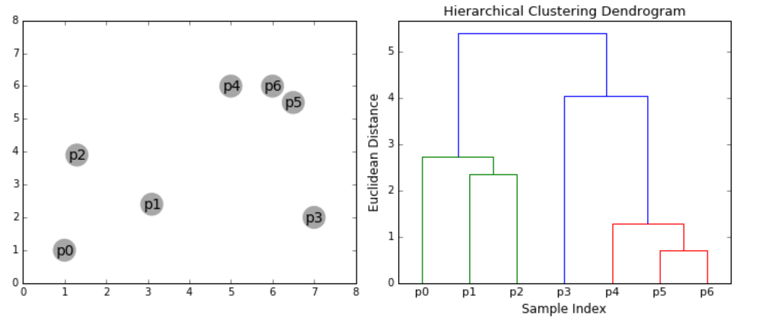
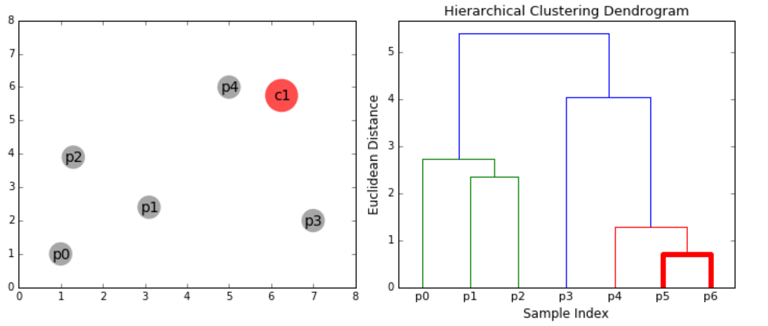
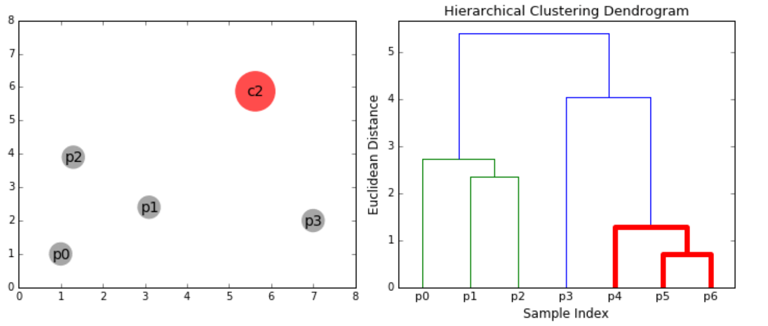
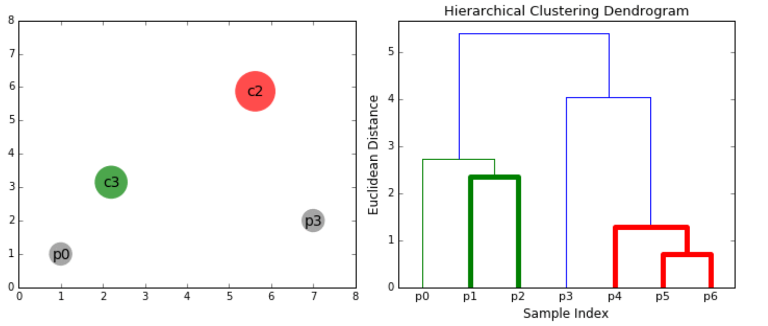
A. Hierarchical Clustering

Hierarchical clustering algorithms fall into 2 categories: top-down or bottom-up. Bottom-up algorithms treat each data point as a single cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters ave been merged into a single cluster that contains all data points. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering or HAC[11]. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample. The Following fig 2 shows illustration of hierarchical clustering.

We begin by treating each data point as a single cluster i.e if there are X data points in our dataset then we have X clusters. We then select a distance metric that measures the distance between two clusters[5]. As an example, we will use *average linkage* which defines the distance between two clusters to be the average distance between data points in the first cluster and data points in the second cluster.

On each iteration, we combine two clusters into one. The two clusters to be combined are selected as those with the smallest average linkage. i.e according to our selected distance metric, these two clusters have the smallest distance between each other and therefore are the most similar and should be combined[8,10].

Step 2 in fig 2 is repeated until we reach the root of the tree i.e we only have one cluster which contains all data points. In this way we can select how many clusters we want in the end, simply by choosing when to stop combining the clusters i.e when we stop building the tree.[9]

1	Consider if there are 7 data points(categories of product) in our dataset then we have 7 clusters	
2	we combine two clusters into one based on distance metric that measures the distance between two clusters	
3	Similarly most similar data points are combined into one	
4	Similarly we combined data point p1 p2 into c3	

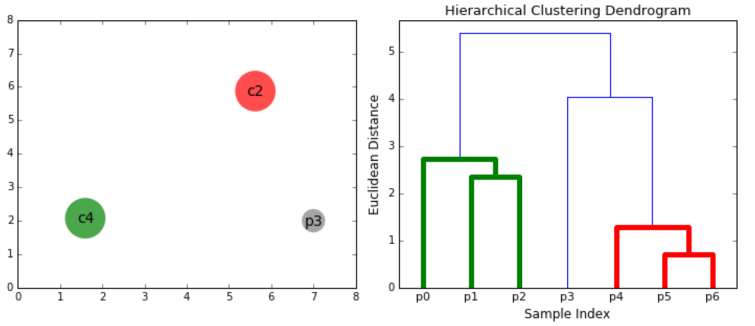
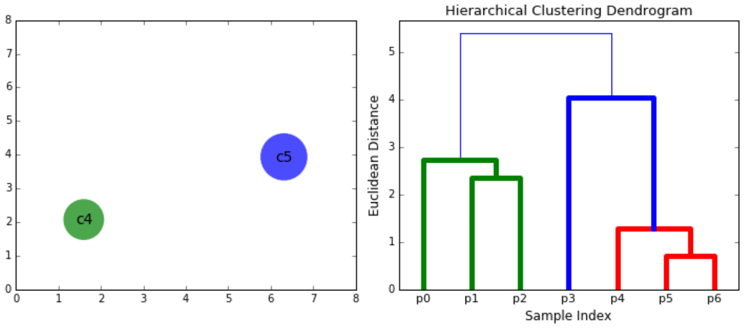
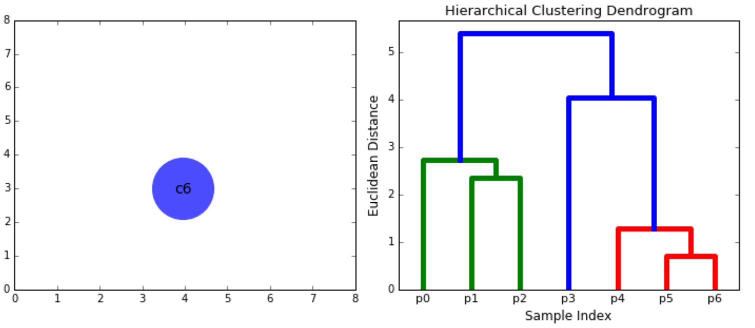
5	These combining process repeated until we reach the root of the tree	
6	Here we combined data point p3 c2 into c5 based on minimum distance between two nodes	
7	At last we reached to major category of all datapoints which represent all product as per our data set	

Fig 2 Hierarchical Clustering implementation

B. *Support Vector Machine*

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.[10,11] A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification, implicitly mapping their inputs into high-dimensional feature spaces. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. Support Vector Machines are based on the concept of decision planes that define decision boundaries[9]. A decision plane is one that separates between a set of objects having different class memberships. Classification tends to, accessing the data sets, through which we are define the classes and sub classes of the data sets.

For example:- if we are using product table, there are several classes which we can define according to the product like Dairy, grooming, bathing, crockeries. After that we classify the class of dairy product into several sub classes i.e. Milk, ghee, butter, butter milk, curd etc[11,12].

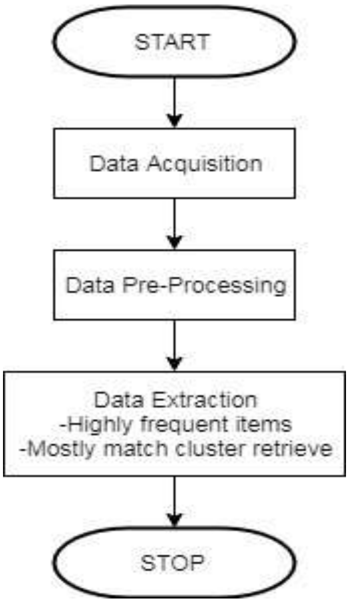


Fig 3 Data extraction using Support Vector Machine

IV. RESULTS

After applying multiple data mining techniques, the results obtained efficient. These tables are quite self explanatory and efficient as per view of user purchasing experience in the sense that they provide the appropriate suggestions raised by system. The K-mean algorithm further support to marketing strategy in the analysis of customer behavior in a way that the products that are closely related together, in terms of use or offered in a deal together are more of a chance to be bought together. Like in the example tea and milk provides confidence of being purchased together.

Following performance graph shows time required to provide 40 items as a suggestions based on selected item of user. Once user buy particular item from list of all items then our mining engine analyses the available data and provide suggestion of item list.

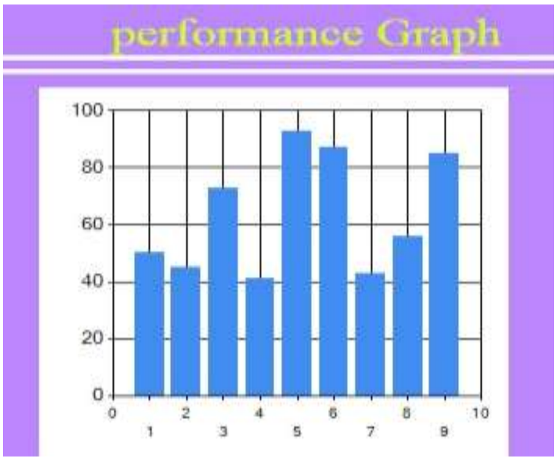


Fig 4 Performance graph

Y axis shows time requires (Mili-seconds) for calculating and providing list of all suggested items and X axis shows different cases of selected inputs. This graph is dynamically generated once user selects particular item for its purchasing

V. CONCLUSIONS

As it is discussed above target of analyzed behavior of such people who are visiting the online shopping sites and spending their time there, checking different stuff. We have taken the database of a related to the shopping malls. Number of products and categories related to the product is present in it. We mined the data from database, ultimate mining algorithms used and queries are applied to extract the data. K means mining and SVM algorithm is applied for the customer behavior analysis.

Support and confidence is the result of the K means mining and SVM algorithm which is implemented on the association of different product, like A→B and B→A, support and confidence is gained from this association.

In future Advance machine learning algorithm can be applied to observe the customer behavior analysis through support and confidence of product mining association rule.

REFERENCES

[1] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE “Data Mining with Big Data” 1041-4347/14.

[2] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). “The Elements of Statistical Learning: Data Mining, Inference and Prediction”. Retrieved 2012-08-07.

[3] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic. “From Data Mining to Knowledge Discovery in Databases”. Retrieved 17 December 2008.

[4] Arun K Pujari, “Data Mining Techniques”, University Press, second edition,2009

[5] Aggarwal, C. C. Charu and C. X. Zhai, Eds., “Chapter 4: A Survey of Text Clustering Algorithms” in Mining Text Data. NewYork: Springer, 2012.

[6] Laurent Galluccioa , Olivier Michelb, Pierre Comonb, Mark Kligerc, Alfred O. Herod, “Clustering with a new distance measure based on a dual-rooted tree”, Information Sciences Volume 251, 1 December 2013, Pages 96-113, Elsevier.

[7] Yiheng Chen and Bing Qin “The Comparison of SOM and K-means of text clustering” School of Computer Science and Technology, Harbin Institute of Technology

[8] “Data Mining Curriculum” ACM SIGKDD. 2006-04-30. Retrieved 2011-10-28.

[9] Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

[10] X. Wu, “Building Intelligent Learning Database Systems,” AI Magazine, vol. 21, no. 3, pp. 61-67, 2000.

[11] Efraim, T.; Jay, E. A.; Tin-Peng, L. & Ramesh, S. (2007). “Decision Support and Business Intelligent Systems, Pearson Education”.

[12] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference and Prediction". Retrieved 2012-08-07.