ISSN 2319-2720

Volume 1, No.2, September – October 2012
International Journal of Computing, Communications and Networking
Available Online at http://warse.org/pdfs/ijccn04122012.pdf

A Study of Various Clustering Algorithms on Retail Sales Data

[1]Vishal Shrivastava, [2]Prem narayan Arya
[1]M.Tech.(Software Systems), SATI, Vidisha, India. shrvishal@gmail.com
[2]Asst. Prof. Dept. of Computer Applications, SATI, Vidisha, India. Premnarayan.arya@rediffmail.com

ABSTRACT

Data mining is the process of extraction of Hidden knowledge from the databases. Clustering is one the important functionality of the data mining Clustering is an adaptive methodology in which objects are grouped together, based on the principle of optimizing the inside class similarity and minimizing the class-class similarity. Various clustering algorithms have been developed resulting in a better performance on datasets for clustering. The paper discusses the four major clustering algorithms: K-Means, Density based, Filtered, Farthest First clustering algorithm and comparing the performances of these principle clustering algorithms on the aspect of correctly class wise cluster building ability of algorithm .The results are tested on datasets of retail sales using WEKA interface and compute the correctly cluster building instances in proportion with incorrectly formed cluster. A comparison of these four algorithms is given on the basis of percentage of incorrectly classified instances.

Keywords: Data mining, Clustering, k means, Retail Sales.

1. INTRODUCTION

The process of Knowledge discovery executes in an iterative sequence of steps such as cleaning of data, its integration, its selection, & transformation of data, data mining, evaluating patterns and presentation of knowledge. Data mining features are characterization and discrimination, mining Frequent patterns, association, correlation, Classification and prediction, cluster analysis, outlier analysis and evolution analysis Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another and very dissimilar to object in other clusters..Dissimilarity is due to the attributes values that describe the objects.

The objects are grouped on the basis of the principle of optimizing the intra-class similarity and reducing the inter-class similarity to the minimum. First of all the set of data is portioned into groups on the basis of data similarity (e g by clustering) and the then assigning labels to the comparatively smaller number of groups.

Several clustering techniques are there: partitioning methods, hierarchical methods, density based methods, grid based methods, model based methods, methods for high dimensional data and constraint based clustering. Clustering is also called data segmentation because clustering partitions large data sets into groups according to their similarity.

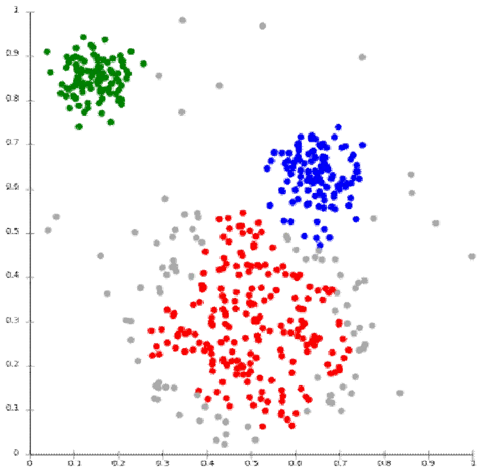

Figure 1: Cluster Analysis

Clustering can be utilized for detecting outlier where outliers are more interesting then common cases e g monitoring of criminal activities in electronic commerce , Credit card fraud detection etc. Clustering is a pre-processing step in the sequence for other algorithms of characterization, attribute subset selection and classification, which then operate on the detected clusters and the selected attributes or features. Research areas include data mining, statistics, machine learning, biology, special database technology and marketing. Clustering is an unsupervised learning. Different from classification, it does not rely on predefined classes and class labels training examples. So clustering is learning by observation and not learning by examples.

A "clustering" is a set of such clusters, that usually contains all objects in the data set. Additionally, it also informs the relationship of the clusters with each other, for example a chain hierarchy of clusters put inside or embedded in each other Clustering can broadly be distinguished into:

Hard clustering: each object belongs to a cluster or no.

Soft clustering (fuzzy clustering): each object belongs to a cluster to a certain degree (like the similarity in belonging to the cluster)

There are also minute distinctions possible, like:

Strict partitioning clustering: Every object belongs to only one cluster

Strict partitioning clustering with outliers: objects can also be of no cluster, and are considered outliers.

Overlapping clustering (also called: alternative clustering, multi- view clustering): objects belong to more than one cluster.

Hierarchical clustering: objects of a child cluster also belong to the parent cluster

Subspace clustering: while an overlapping clustering, clusters are not expected to overlap.

Desired Typical requirements of clustering in data mining are
(i) Scalability (ii) ability to deal with various types of attributes, (iii) Discovery of clusters with different shapes,(iv) Minimal requirement for domain knowledge to determine input parameters, (v)Ability to deal with noisy data,(vi) Incremental Clustering and insensitivity to the order of input records, (vii)High dimensionality,(viii) Constraint based Clustering and Interpretability and (ix) usability.

k-Means ClusterinG

We determine number of clusters N and we assume the centroid of these clusters. We can take any random objects as the initial centroids or the first N objects that can also serve as the initial centroids. Then the N means algorithm will perform the three steps given below until convergence occurs. Iterate until stability (= no object move group):

1. Determine the coordinate of centroid
2. Determine the distance of each object from the centroids
3. Unite the Group of the objects based on minimum distance (finding the closest centroid simultaneously). This is showed in figure 1.1 in steps.

Density based Clustering

To discover clusters with arbitrary shape, density based clustering methodology have been developed hence typically regard clusters as dense region of objects in the data space that are separated by regions of low density.

Filtered Clustering

A filter adds a new nominal attribute that represents the clusters assigned to every instance by specified clustering algorithm. Either the clustering algorithm is built with the first batch of data or ones specifications are serialized clustered model file to use, instead.

Farthest First Clustering

Farthest first is a variant of N Means that places all the cluster centre in turn at the point which is farthest from the existing cluster centre. This point should be within the data area. This speed up the clustering in most of the cases greatly since lesser reassignments and adjustments are needed.

2. REVIEW OF LITERATURE

Xiaozhe Wang et al., in 2006 provided a method of clustering of the time series based on their structural characteristics, rather it groups based on global features extracted from the time series. Global measures explaining the time series are achieved by applying statistical methods that capture the following characteristics: trend, nonlinearity, skewness, seasonality, chaos, , periodicity, serial correlation, kurtosis, and self-similarity. Since the method clusters use extracted global measures, it minimizes the dimensionality of time series and is very less sensitive to noisy data. A search mechanism is then provided to find the best selection from the feature set that should be used for the clustering inputs [2].

Li Wei et al. in 2005 defined a tool for visualizing and data mining of medical time series and found that increasing the interest in time series data mining has had astonishingly minute impact on real world medical applications. Practitioners working with time series regularly, rarely take advantage of the tools that the data mining community has made available. This approach finds features from a time series of random length and utilizes information about the relative frequency of these features to color an image in a principled way. By observing the similarities and differences within a collection of image bitmaps, a user can quickly find clusters, exceptions, and other regularities within the data collection.

An Online Algorithm for Segmenting Time series was executed by Eamonn Keogh et al., in 2001. This was the first wide review and formulated comparison of time series segmentation algorithms from a data mining point of view. Thus emerged the most popular approach, Sliding Windows, which generally produces poor results, and the second most popular approach, Top-Down, will produce reasonable results, it is not scalable. On the contrary, least known, Bottom-Up approach will produces excellent results and it scales linearly with the size of the dataset. In addition, this introduced SWAB, a new algorithm, which also scales linearly with the size of the dataset, and requires only constant space and produces high quality approximations of the data.

A Model Based Clustering for Time Series with Irregular Interval was proposed by Xiao-Tao Zhang et al., in 2004. This focussed Clustering problems are central to many knowledge discovery and data mining tasks. However, most existing clustering methods can only work with fixed interval representations of data patterns, ignoring the variance of time axis. This studied the clustering of data patterns that are sample in irregular interval. A model-based approach.This focused on the clustering of data patterns which are sampled in irregular interval. A model-based approach that use cepstnun distance metric and Autoregressive Conditional Duration (ACD) model has proposed. Experimented results on real datasets endorses that this method is effective in clustering irregular space time series, and also results inferred from experimental values agrees with the market microstructure theories.

Hui Ding et al., in 2008 used experiments to compare the representations and distance measures of querying and mining of Time Series Data. This led to conducting an extensive set of time series experiments re-implementing 8 different representation methods and 9 similarity measures and their variants, and testing their effectiveness on 38 time series data sets from a wide variety of application perspectives. They provided an outline of these techniques and presented their comparative experimental results corresponding to their effectiveness. Their experiments have provided both a unified validation of existing achievements, and suggested that certain claims in the literature may be hopeful.

Ehsan Hajizadeh et al.,in 2010examined and provided an outline of application of data mining like decision trees, association rules, neural network, factor analysis and etc in the stock markets. Also, this tells progressive applications known gap and less significant area and determined the future works for researchers. This tells the problems of data mining in finance (stock market) and specific requirements for data mining methods including in making Interpretations ,incorporating relations and probabilistic learning. The data mining techniques mentioned here increases the performance in pattern discovery methods that deals with rigorous numeric and alpha numeric data, that involves structured objects, text and data in a variety of discontinuous and continuous scales (nominal, order, absolute and so on).

Also, this show benefits of using such techniques for stock market forecast[7]. Jiangjiao Duan et al., in 2005 introduced that Model-based clustering is one of the most important ways for time series data mining. However, the process of clustering may encounter several problems. Here a novel clustering of the dataset. In addition, this introduced SWAB, a new online algorithm, which scales linearly with the size of the dataset, requires only constant space and produces high quality approximations of the data [4].

Jiangjiao Duan et al., in 2005 introduced that Model-based clustering which is one of the important ways for time series data mining. However, the process of clustering may face several problems. Here a novel clustering algo of time-series incorporating recursive Hidden Markov Model (HMM) training was proposed. It contributed in following aspects:
1) It recursively trains models and use this model information in the process agglomerative hierarchical clustering.
2) It built HMM of time series clusters to describe clusters.

To evaluate the effectiveness of the algorithm, so many experiments have been conducted on both synthetic and real world data. The inferences shows that this approach can achieve nicer performance in correctness rate than the conventional HMMbased clustering algorithm [8]. Information Mining Over Heterogeneous and High-Dimensional Time-Series Data in Clinical Trials Databases

was carried out by Fatih Altiparmak et al., in 2006. They gave a good approach for data mining involving two steps:
(i) applying a data mining algorithm over homogeneous subsets of data, and identifying common or (ii) distinct patterns over the information gathered in the first step. This approach is implemented only for different and high dimensional time series clinical trials of data. Using the framework, they propose a new way of utilizing frequent item set mining, as well as clustering and declustering techniques with better distance metrics for measuring similarity among time series data. By grouping the data, it find groups of analyze (substances present in blood) which is most strongly correlated. Most of these known relationships are verified by the clinical panels, and, in addition, they identify novel groups that require further biomedical analysis. A slight change in the method results in an effective declustering of high dimensional time series data, which is then used in "feature selection." Using industry-sponsored clinical trials data sets, they are able to find out a smaller set of analytes that effectively models the status of normal health [9].

3. COMPARATIVE PERFORMANCE ANALYSIS OF CLUSTERING ALGORITHMS

For performance evaluation of the four most popular clustering techniques K-Mean clustering, Density based Clustering, Filtered clustering and farthest first clustering, we have taken datasets containing numerical attribute then convert these into nominal attributes type that is all these datasets contains the continuous attributes. This dataset contain 26 attribute on the basis of assumption class attribute as No. of bays, the cluster are generating by applying the below mentioned algorithms using the Weka interface. Weka is a landmark system in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time (the first version of Weka was released 11 years ago).

We have performed various experiment on the retail sales data. We have performed following 4 clustering method on the data.

i. Simple kMeans Clustring
ii.Density Based Clustring
iii.Filtered Clustring
iv.Farthest First Clustring

We have already discussed various clustring methods, in previous section. All the above clustering methods are applied on the data detailed as in the next section. These methods are applied with the help of weka data mining tool & take the results.

Retail sales is an important issue in the market. Various research are done in this area to understand the market pattern, customer behaviour, demand & supply, inventory etc. in this thesis we also used retaile sales data. The original data contain 126 attribute & 3067 intances but in our experiments we take the following 26 attribute with all 3067 instances. This is the sales per capita data of different states of america.

Table 1: All Variable Names

S.No.	Variable Name	Definition
1	STATE	State Identifier ICPSR
2	PCTBLK3	% black in 1930
3	PCTURB3	% urban in 1930
4	PCTFRM3	% of land on farms in 1929
5	PCTILL3	% illiterate over age 10, 1930
6	PFORB3	% foreign born
7	NDMTCODE	County code based on ICPSR county code with adjusted values for combined counties
8	AREA	area in square miles
9	LATITUDE	Latitude of county seat
10	LONGITUD	Longitude of county seat
11	STAB	State Name
12	CNAMESR	County Name
13	PP301019	% population aged 10-19, 1930
14	PP302029	% population aged 20-29, 1930
15	PP303034	% population aged 30-34, 1930
16	PP303544	% population aged 35-44, 1930
17	PP304554	% population aged 45-54, 1930
18	PP305564	% population aged 55-64, 1930
19	PP3065UP	% population aged 65 and over, 1930
20	BAY	Number of Bays in County
21	BEACH	Number of Beaches
22	LAKE	Number of Lakes
23	RRTSAP39	Retail Sales per Capita in 1939 in 1967 $
24	RRTSAP35	Retail Sales per Capita in 1935 in 1967 $

25	RRTSAP33	Retail Sales per Capita in 1933 in 1967 $
26	RRTSAP29	Retail Sales per Capita in 1929 in 1967 $

4.. EXPERIMENTAL SIMULATION AND RESULTS

Above four algorithms have their implemented source code in the Weka upon which simulations have carried out in order to measure the performance parameters of the algorithms over the datasets. The results are summarized in the following tables & Graph.

Table 2: Cluster Distribution

Name of the Cluster Method	No. of cluster	Instance in cluster 1	Instance in cluster 2
Simple kMeans	2	1675	1392
Density based Cluster	2	1688	1379
Filtered Cluster	2	1675	1392
Farthest First	2	2904	163

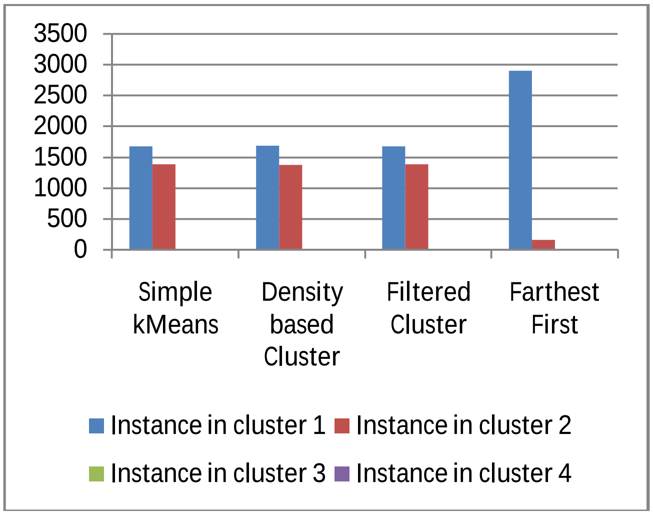

Figure 2: Cluster Distribution

Table 2: Clustering Performance

Method	No of cluster	Incorrectly Clustered instances	% of Incorrectly Clustered instances
Simple kmeans	4	1500	48.9077
Simple kmeans	3	1436	46.821
Simple kmeans	2	964	31.4314
Density Based Cluster	2	965	31.464
Filtered Cluster	2	964	31.4314
Farthest First	2	804	26.2145
Farthest First	3	816	26.6058
Farthest First	4	814	26.5406

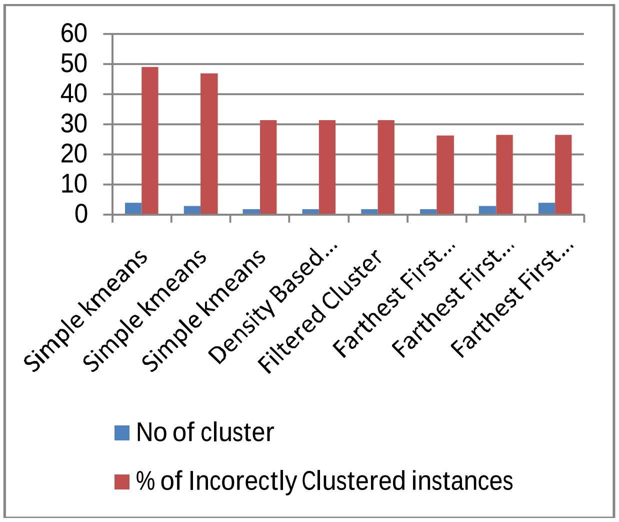

Figure 3: Clustering Performance

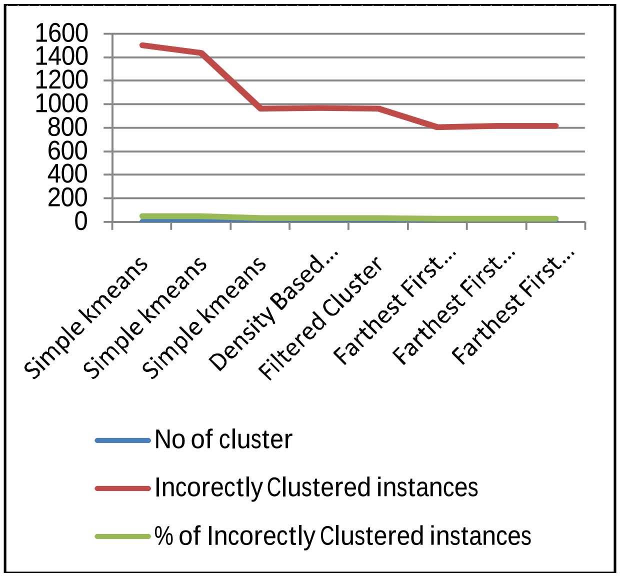

Figure 4: Clustering Performance

5. CONCLUSION

Performance of the clustering method is measured by the percentage of the incorrectly classified instances. As the percentage of the incorrectly classified attribute is low performance of the clustering is as good. Farthest first clustering gives better performance compared to k means clustering, Density based clustering & filtered clustering. Also this algorithm's result is independent of number of cluster while k means algorithm result is highly dependent on the number of cluster. Farthest first clustering though gives a fast analysis when taken an account of time domain, but makes comparatively high error rate. We can see farthest first algorithm give lowest 26.2145 % of incorrectly classified instances.

REFERENCES

1. Han J. and Kamber M. **Data Mining: Concepts and Techniques**, *Morgan Kaufmann Publishers*, San Francisco, 2000.

2. Xiaozhe Wang, Kate Smith and Rob Hyndman. **Characteristic-Based Clustering for Time Series Data**, *Data Mining and Knowledge Discovery, Springer Science + Business Media, LLC Manufactured in the United States,* pp. 335–364, 2006.

3. Li Wei, Nitin Kumar, Venkata Lolla and Helga Van Herle. **A practical tool for visualizing and data mining medical time series**, Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05), pp. 106- 125, 2005.

4. Eamonn Keogh, Selina Chu,David Hart and Michael Pazzani. **An online algorithm for segmenting time series**, 0-7695-1 119-8/01 IEEE, 2001.

5. Xiao-Tao Zhang, Wei Zhang and Xiong Xiong. **A model based clustering for time-series with irregular interval**, Proceedings of the Third International Conference on Machine Learhg and Cybernetics, Shanghai, pp.26-29, August 2004.

6. Hui Ding, Goce Trajcevski and Eamonn Keogh. **Querying and mining of time series data: Experimental comparison of representations and distance measures**, *PVLDB '08*, August, pp. 23-28, 2008, Auckland, New Zealand, 2008.

7. Ehsan Hajizadeh, Hamed Davari Ardakani and Jamal Shahrabi. **Appilication of data mining techniques in stock market**, Journal of Economics and International Finance Vol. 2(7), pp. 109-118, July 2010.

8.Jiangjiao Duan, WeiWang , Bing Liu and Baile Shi. **Incorporating with recursive model training in time series clustering**, Proceedings of the 2005 The Fifth International Conference on Computer and Information Technology (CIT'05), IEEE2005.

9. Fatih Altiparmak, Hakan Ferhatosmanoglu, Selnur Erdal, and Donald C. TrostFaith Altipar. **Information Mining Over Heterogeneous and High-Dimensional Time-Series Data in Clinical Trials Databases**, IEEE Transactions On Information Technology In BioMedicine, Vol.10, pp.215-239, April 2006.

10. Jinfei Xie and Wei-Yong Yan. **A Qualitative Feature Extraction Method for Time Series Analysis**, Proceedings of the 25th Chinese Control Conference, pp. 7–11, August, 2006, Harbin, Heilongjiang, 2006.

11. Xiaoming lin, Yuchang Lu and Chunyi Shi. **Cluster Time Series Based on Partial Information**, IEEE SMC TPUl, pp. 254-262, 2002.

12. Yuan F, Meng Z. H, Zhang H. X and Dong C. R. **A New Algorithm to Get the Initial Centroids**, Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26–29, August 2004.

13. Sun Jigui, Liu Jie and Zhao Lianyu. **Clustering algorithms Research,** Journal of Software ,Vol 19,No 1, pp.48- 61,January 2008.

14. Sun Shibao and Qin Keyun. **Research on Modified kmeans Data Cluster Algorithm,** I. S. Jacobs and C. P. Bean, *Fine particles, thin films and exchange anisotropy*, Computer Engineering, Vol.33, No.13, pp.200–201,July 2007.

15. Merz C and Murphy P. **UCI Repository of Machine Learning databases**, Available:ftp://ftp.ics.uci.edu /pub/machinelearning- databases

16. Fahim A M,Salem A M and Torkey F A. **An efficient enhanced k-means clustering algorithm**, Journal of Zhejiang University Science A, Vol.10, pp:1626-1633,July 2006.