A Comparative Performance Analysis of Clustering Algorithms

Pallavi #, Sunila Godara *

#(Student ,Department of computer science & engg. ,GJU S&T, Hisar)
** (Assistant professor, Department of computer science & engg ., GJU S&T, Hisar)*

ABSTRACT

Clustering is an adaptive procedure in which objects are clustered or grouped together, based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. Various clustering algorithms have been developed which results to a good performance on datasets for cluster formation. This paper analyze the three major clustering algorithms: K-Means, Farthest First and Hierarchical clustering algorithm and compare the performance of these three major clustering algorithms on the aspect of correctly class wise cluster building ability of algorithm. The results are tested on three datasets namely Wine, Haberman and Iris dataset using WEKA interface and compute the correctly cluster building instances in proportion with incorrectly formed cluster. A performance percentage has been calculated to analyze the algorithms on dataset taking Principal Component Analysis as another parameter of comparison.

Keywords – **Clustering, Data mining, Weka.**

I. INTRODUCTION

Knowledge discovery process consists of an iterative sequence of steps such as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. Data mining functionalities are characterization and discrimination, mining frequent patterns, association, correlation, classification and prediction, cluster analysis, outlier analysis and evolution analysis [1]. 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 based on the attributes values describing the objects. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. Firstly the set of data is portioned into groups based on data similarity (e g Using clustering) and the then assign labels to the relatively small 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. Clustering can be used for outlier detection where outliers may be more interesting then common cases e g Credit card fraud detection, monitoring of criminal activities in electronic commerce. Clustering is a pre-processing step for other algorithms such as characterization, attribute subset selection and classification, which would then operate on the detected clusters and the selected attributes or features.

Contributing area of research include data mining, statistics, machine learning, special database technology, biology and marketing. Clustering is an unsupervised learning. Unlike classification, it does not rely on predefined classes and class labels training examples. Hence we say clustering is learning by observation rather than learning by examples.

Typical requirements of clustering in data mining are Scalability, ability to deal with different types of attributes, Discovery of clusters with different shapes, Minimal requirement for domain knowledge to determine input parameters, Ability to deal with noisy data, Incremental Clustering and insensitivity to the order of input records, High dimensionality, Constraint based Clustering and Interpretability and usability.

K-MEANS CLUSTERING: The basic step of k-means clustering is simple. In the beginning, we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects can also serve as the initial centroids. Then the K means algorithm will do the three steps below until convergence.Iterate until *stable* (= no object move group):

1. Determine the centroid coordinate
2. Determine the distance of each object to the centroids
3. Group the object based on minimum distance (find the closest centroid). This is showed in figure 1.1 in steps.

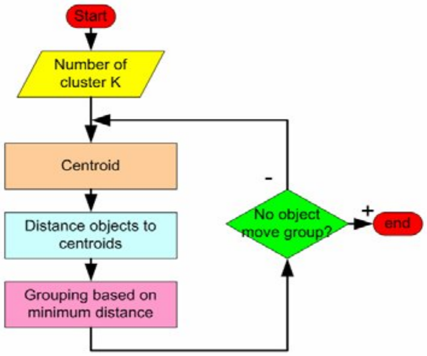
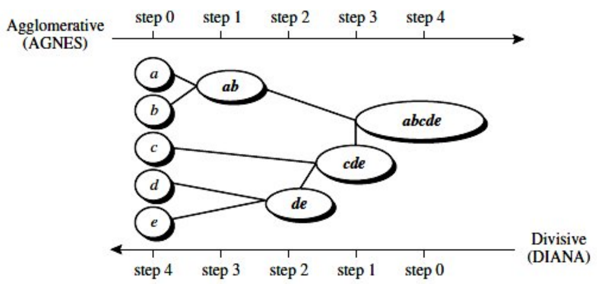

Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

Figure 1.1: k-means clustering process

HIERARCHICAL CLUSTERING: Hierarchical Clustering Methods are Agglomerative hierarchical methods. This Begins with as many clusters as objects. Clusters are successively merged until only one cluster remains. Divisive hierarchical methods begin with all objects in one cluster. Groups are continually divided until there are as many clusters as objects.



Agglomerative and divisive hierarchical clustering on data objects $\{a, b, c, d, e\}$.

Figure 1.2: Hierarchical Clustering Process

FARTHEST FIRST CLUSTERING: Farthest first is a varient of K Means that places each cluster centre in turn at the point furthermost from the existing cluster centre. This point must lie within the data area. This greatly speeds up the clustering in most of the cases since less reassignment and adjustment is needed.

II. REVIEW OF LITERATURE

Xiaozhe Wang et al., in 2006 provide a method for clustering of time series based on their structural characteristics, rather it clusters based on global features extracted from the time series. Global measures describing the time series are obtained by applying statistical operations that best capture the underlying characteristics: trend, seasonality, periodicity, serial correlation, skewness, kurtosis, chaos, nonlinearity, and self-similarity. Since the method clusters using extracted global measures, it reduces the dimensionality of the time series and is much less sensitive to missing or noisy data. A search mechanism is provided to find the best selection from the feature set that should be used as the clustering inputs [2].

Li Wei et al. in 2005 stated a practical tool for visualizing and data mining medical time series and concluded that increasing interest in time series data mining has had surprisingly little impact on real world medical applications. Practitioners who work with time series on a daily basis rarely take advantage of the wealth of tools that the data mining community has made available. This approach extracts features from a time series of arbitrary length and uses information about the relative frequency of these features to colour a bitmap in a principled way. By visualizing the similarities and differences within a collection of bitmaps, a user can quickly discover clusters, anomalies, and other regularities within the data collection [3].

An Online Algorithm for Segmenting Time series was carried out by Eamonn Keogh et al., in 2001. This was the first extensive review and empirical comparison of time series segmentation algorithms from a data mining perspective. This showed the most popular approach, Sliding Windows, generally produces very poor results, and that while the second most popular approach, Top-Down, can produce reasonable results, it does not scale well. In contrast, the least well known, Bottom-Up approach produces excellent results and scales linearly with the size 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].

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 using cepstnun distance metric and Autoregressive Conditional Duration (ACD) model has proposed. Experimental results on real datasets showed that this method is generally effective in clustering irregular space time series, and conclusion inferred from experimental results agrees with the market microstructure theories. [5]

Hui Ding et al., in 2008 experimentally compared the representations and distance measures of querying and mining of Time Series Data. This conducted 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 domains. They gave an overview of these different techniques and present their comparative experimental findings regarding their effectiveness. Their experiments have provided both a unified validation of some of the existing achievements, and in some cases, suggested that certain claims in the literature may be unduly optimistic [6].

Ehsan Hajizadeh et al.,in 2010 provided an overview of application of data mining techniques such as decision tree, neural network, association rules, factor analysis and etc in stock markets. Also, this reveals progressive applications in addition to existing gap and less considered area and determines the future works for researchers. This stated 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 outlined here advances pattern discovery methods that deals with complex numeric and non-numeric data,

involving structured objects, text and data in a variety of discrete 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 algorithm of time-series which incorporates recursive Hidden Markov Model (HMM) training was proposed. It contributed the following aspects: 1) It recursively train 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, several experiments had conducted on both synthetic data and real world data. The result shows that this approach can achieve better performance in correctness rate than the traditional HMM-based 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 novel approach for information mining that involves two major steps: applying a data mining algorithm over homogeneous subsets of data, and identifying common or distinct patterns over the information gathered in the first step. This approach implemented specifically for heterogeneous and high dimensional time series clinical trials data. Using this framework, this propose a new way of utilizing frequent item set mining, as well as clustering and declustering techniques with novel distance metrics for measuring similarity between time series data. By clustering the data, it find groups of analyze (substances in blood) that are most strongly correlated. Most of these relationships already known are verified by the clinical panels, and, in addition, they identify novel groups that need further biomedical analysis. A slight modification to this algorithm results an effective declustering of high dimensional time series data, which is then used for “feature selection.” Using industry-sponsored clinical trials data sets, they are able to identify a small set of analytes that effectively models the state of normal health [9].

III. COMPARATIVE PERFORMANCE ANALYSIS OF CLUSTERING ALGORITHMS

For performance evaluation of the tree most popular clustering techniques K-Mean clustering, Hierarchical clustering and Farthest first clustering, we have taken three datasets containing nominal attributes type that is all these datasets contains the continuous attributes. Each dataset’s instance has contained an assigned class with it. On the basis of this class the cluster are generating by applying the above 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). These datasets have been taken from UCI machine learning repository system.

Iris plants dataset contains 3 classes of 50 instances each where each class refers to a type of iris plant. One class is linearly separable from the other 2, the latter are NOT linearly separable from each other. No. of instances are 150(50 in each of the 3 classes). No of attributes are 5 including the class attributes.

Haberman’s Survival Dataset contains cases from a study that was conducted on the survival of patients who had undergone surgery of breast cancer. No. of instances 306 and no. of attributes are 4 including the class attribute.

Wine Recognition Dataset shows the result a chemical analysis of wines grown in the same region in Italy but derived from 3 different cultivators. The analysis determined the quantities of 13 constituents found in each of the 3 types of wines. No. of instances are 178 and no. of attributes are 13.These datasets taken in to arff format for being able to process on the Weka interface. Then one by one these datasets are evaluated and their clustering performance is evaluated by evolving the parameter without principal component analysis and with including the component analysis. Amount of correctly clustered instances and incorrectly clustered instances have been recorded and the root mean square error is computed for each clustering algorithm. Each algorithm is run over three predefined datasets and their performance is evaluated.

IV. EXPERIMENTAL SIMULATION AND RESULTS

These three algorithms have their implemented source code in the Weka 3.6.4 version 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.

TABLE 1: K-MEANS CLUSTERING PERFORMANCE WITHOUT PRINCIPLE COMPONENT ANALYSIS FILTER

Dataset	No of class	Incorrect cluster instances	% Error	No of iterations	Squared error
Iris	3	18	12%	4	7.13
Haberman	2	148	18%	7	254.3
Wine	3	9	5.05%	5	48.98

TABLE 2: HIERARCHICAL CLUSTERING PERFORMANCE WITHOUT PRINCIPLE COMPONENT ANALYSIS FILTER

Dataset	Incorrect cluster instansces	Percentage % Incorrect instance
Iris	51	34%
Haber man	90	29.4%
Wine	109	61.2%

TABLE 3: FARTHEST FIRST CLUSTERING PERFORMANCE WITHOUT PRINCIPLE COMPONENT ANALYSIS FILTER

Dataset	Incorrectly cluster instances	Percentage% Incorrect instance
Iris	69	46%
Haber man	135	44.11%
Wine	66	37.07%

TABLE 4: K-MEANS CLUSTERING PERFORMANCE WITH PRINCIPLE COMPONENT ANALYSIS FILTER

Dataset	Incorrect cluster instances	Percentage% Incorrectly clusters	No of iterations	Squared error
Iris	27	18%	5	3.74
Haber man	102	33.3%	3	188.3
Wine	9	5.0%	5	48.9

TABLE 5: HIERARCHICAL CLUSTERING PERFORMANCE WITH PRINCIPLE COMPONENT ANALYSIS FILTER

Dataset	Incorrect cluster instansces	Percentage% of incorrectly Clustered instance
Iris	51	34%
Haber man	88	28.7%
Wine	109	61.2%

TABLE 6: FARTHEST FIRST CLUSTERING PERFORMANCE WITH PRINCIPLE COMPONENT ANALYSIS FILTER

Dataset	Incorrectly cluster instances	Percentage % Incorrectly clustered
Iris	51	34%
Haber man	120	39.21%
wine	66	37.07%

Two successive iterations have been taken for each algorithm on each dataset. First a non-filter approach has taken without inserting the principle component analysis filter and second time with inserting the principle component analysis filter.

V. CONCLUSION

The result analysis shows that K-means algorithm performs well without inserting the principle component analysis filter as compared to the Hierarchical clustering algorithm and Farthest first clustering since it have less instances of incorrectly clustered objects on the basis of class clustering. Hierarchical clustering as compared to Farthest fast clustering gives better performance. Also this algorithm performs better after merging principle component analysis filter with it. Farthest first clustering though gives a fast analysis when taken an account of time domain, but makes comparatively high error rate. Using principle component analysis filter with this approach, this shows better results which are comparable with Hierarchical clustering algorithm.

VI. 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*, 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)* 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*, 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 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, 215-239, APRIL 2006.

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

[11] Xiaoming lin, Yuchang Lu, Chunyi Shi:” Cluster Time Series Based on Partial Information”, *IEEE SMC TPUl*, p. no. 254-262, year 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, Zhao Lianyu, “Clustering algorithms Research”,Journal of Software ,Vol 19,No 1, pp.48-61,January 2008.

[14] Sun Shibao, Qin Keyun,”Research on Modified k-means 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/machine-learning-databases

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