Volume 2, Issue 5, May 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Comparative Performance Analysis of Clustering Algorithms

Shreya Jain* *Computer Science & Engg* *India*
Samta Gajbhiye *Computer Science & Engg* India

Abstract— **Cluster analysis is important for analyzing the number of clusters of natural data in several domains. Various clustering methods have been proposed. However, it is very difficult to choose the method best suited to the type of data. Therefore, the objective of this research was to compare the effectiveness of four clustering techniques with multivariate data. The techniques were:**
K means Clustering algorithm, Fuzzy C Means algorithm, competitive Neural Network and one novel method Quantum Clustering was added to evaluate its relative performance. Such algorithms may employ distinct principles, and lead to different performance and results. The appropriate choice of a clustering method is a significant and often overlooked aspect in extracting information from large-scale datasets. Evidently, such choice may significantly influence the biological interpretation of the data. We present an easy-to-use and intuitive tool that compares some clustering methods within the same framework. It first reduces the dataset's dimensionality using the Singular Value Decomposition (SVD) method, and only then employs various clustering techniques. Besides its simplicity, and its ability to perform well on high dimensional data, it provides visualization tools for evaluating the results. We tested various algorithms on a variety of datasets.

Keywords— **K-Means, Fuzzy C- Means, Competitive Net, Quantum Clustering, Jaccard Score.**

I. INTRODUCTION

In the field of genomics and proteomics, as well as in many other disciplines, classification is a fundamental challenge. Classification is defined as systematically arranging entities (data-points) into specific groups. Clustering, being an unsupervised learning problem, may be regarded as a special case of classification with unknown labels. Statistical methods are used in current research in several domains including: social sciences, management, medicine, agriculture and other sciences. Almost all research needs to collect large amounts of data and manage it systematically in order to analyze processes or systems. Data clustering is one of the important analytical techniques and will become increasingly useful in the future, for visualizing data and searching for hidden trends in the data. Cluster analysis is a class of statistical technique used to separate data into appropriate groups. It is most important in unsupervised learning problems since these techniques deal with finding structure in a collection of unlabeled data. Clustering algorithms can be divided into two types: hierarchical algorithms and partitional algorithms.
Hierarchical algorithms, such as hierarchical clustering, begin with matching each object with similar ones that are placed in a separate cluster and then merged into larger clusters. On the other hand, partitional algorithms, such as K-means clustering, classify the whole object into smaller clusters. Many researchers have proposed other clustering techniques for various data.

Kaufman and Rousseeuw (1990) considered it can be a challenging problem to choose the best clustering algorithm from the many available. Therefore, the main purpose of this work was to compare the effectiveness of four techniques:
K means Clustering algorithm, Fuzzy C Means algorithm, competitive Neural Network and Quantum Clustering.

In this paper for a comparative analysis we included routinely used clustering algorithms and commonly applied statistical tests, such as K-Means, Fuzzy C-Means and a competitive neural network and Quantum Clustering (QC). We conclude that the compression of data that comprises the first step in the tool not only reduces computational complexity but also improves clustering quality. Interestingly, in the presented tested datasets the QC algorithm outperforms the others.

This report is organized as follows : following this introduction, it is included in Section 2 describes the brief review of the literature, Section 3 describes the Implementation procedure, Section 4 describes the description of various algorithms which is used in this software, Section 5 describes the Comparative Analysis of Various Clustering algorithms, Section 6 describes the Experimental Simulation and Results and finally Section 7 provides conclusions and describes directions for future research.

II. LITERATURE REVIEW

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].

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 [4].

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 [5].

III. IMPLEMENTATION

In the presented software, four steps should be followed: defining input parameters, pre-processing, selecting the clustering method and presenting the results.

A. *Input Parameters*

The software receives two input parameters that are Matlab variables: data (a two dimensional matrix) – represents the elements to be clustered, and 'real classification (an optional, one-dimensional vector) – representing the elements according to an expert view and is based on bulk biological and medical knowledge.

B. Pre-processing

a) Determining the matrix shape and which vectors are to be clustered (rows or columns).
b) Pre-processing Procedures: SVD, normalization and dimension selection.

C. Selecting the Clustering Method

a) Points' distribution preview and clustering method selection: The elements of the data matrix are plotted. If a 'real classification' exists, each of its classes is displayed in a different colour. One of the clustering methods, K-means, FCM
(Fuzzy C-means), Competitive NN (Neural Network) or QC (Quantum Clustering) is to be chosen from the menu.
b) Parameters for clustering algorithms: depending on the chosen method, a specific set of parameters should be defined (e.g., in the K-Means method – number of clusters).

D. Results

Once the software completes its run, the results are displayed in both graphical and textual formats (results can be displayed also in a log window). In the graphical display, points are tagged by the algorithm. The textual display represents Purity and Efficiency (also known as precision and recall or specificity and sensitivity, respectively) as well as the joint Jaccard Score.

$$\text{Purity} = n_{11} / n_{11} + n_{01}$$

$$\text{Efficiency} = n_{11} / n_{11} + n_{10}$$

$$\text{Jaccard} = n_{11} / n_{11} + n_{01} + n_{10}$$

Where:
• n_{11} is the number of pairs that are classified together, both in the 'real' classification and in the classification obtained by the algorithm.
• n_{10} is the number of pairs that are classified together in the correct classification, but not in the algorithm's classification.
• n_{01} is the number of pairs that are classified together in the algorithm's classification, but not in the correct classification.

Ending the application will add a new variable to the Matlab workspace: calcMapping- a one-dimensional vector that represents the calculated classification of the elements.

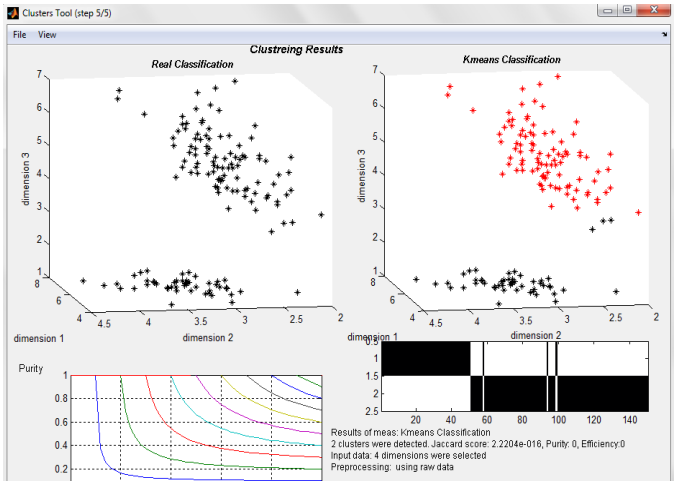


Fig 1 A screenshot of the graphical view of the results

IV. DESCRIPTION OF VARIOUS ALGORITHMS

A. *K-Means Clustering*

The K-means algorithm is one of the partitioning based, non-hierarchical clustering methods. Given a set of numeric objects ***X*** and an integer number *k*, the K-means algorithm searches for a partition of ***X*** into *k* clusters that minimizes the within groups sum of squared errors. The K-means algorithm starts by initializing the *k* cluster centres. The input data points are then allocated to one of the existing clusters according to the square of the Euclidean distance from the clusters, choosing the closest. The mean (centroid) of each cluster is then computed so as to update the cluster centre. This update occurs as a result of the change in the membership of each cluster. The processes of re-assigning the input vectors and the update of the cluster centres is repeated until no more change in the value of any of the cluster centres.

B. *Fuzzy C-Means Clustering*

The Traditional Fuzzy C Means algorithm is one of the most widely used fuzzy clustering algorithms. This technique was originally introduced by Jim Bezdek in 1981.
The Fuzzy C Means algorithm attempts to partition a finite collection of elements X={ x1, x2, ..., xn} into a collection of C fuzzy clusters with respect to some given criterion. Fuzzy sets allow for degrees of membership. A single point can have partial membership in more than one class. There can be no empty classes and no class that contains no data points. The output of such algorithms is a clustering, but not a partition some times. Fuzzy clustering is a widely applied method for obtaining fuzzy models from data. It has been applied successfully in various fields including census, surveying, finance, earthquakes or marketing. This method is frequently used in pattern recognition.

C. *Competitive Neural Network*

Competitive learning can be implemented using a two-layer (J-K) neural network, as shown in Fig. 2. The input and output layers are fully connected. The output layer is called the competition layer, wherein lateral connections are used to perform lateral inhibition.

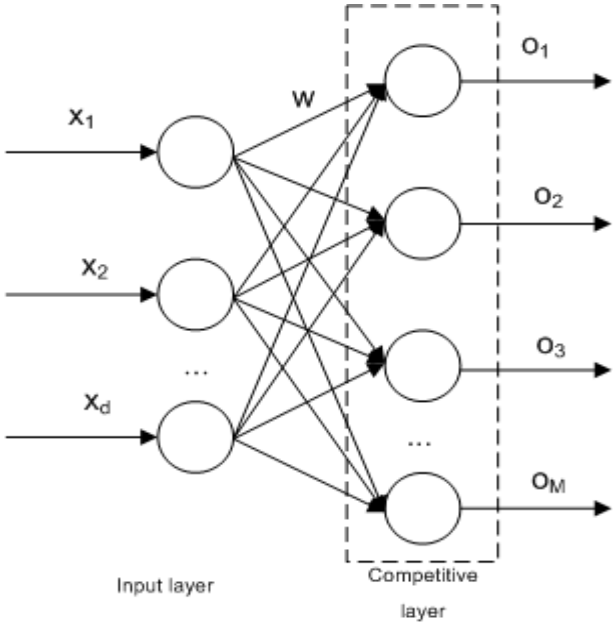


Fig 2 Architecture of Competitive Learning Network

Based on the mathematical statistics problem called cluster analysis, competitive learning is usually derived by minimizing the mean squared error (MSE) functional (Tsypkin, 1973)

$$E = 1/N \sum_{p=1}^{N} E_p$$

$$E_p = \sum_{k=1}^{K} \mu_{kp} \, \|x_p - c_k\|^2$$

where N is the size of the pattern set, and μ_{kp} is the connection weight assigned to prototype c_k with respect to x_p, denoting the membership of pattern p into cluster k. When c_k is the closest (winning) prototype to x_p in the Euclidean metric, $\mu_{kp} = 1$; otherwise $\mu_{kp} = 0$. [6]

D. *Quantum Clustering*

A novel method of clustering by Horn et al. [7], called Quantum Clustering, was also implemented.
This physics-inspired method creates a probabilistic wave function and potential function, which constitute a solution to the Schrodinger equation. Singular value decomposition is first performed, then each data point is assigned to a Gaussian

of width σ by a Parzen-window approach. One free parameter is varied and indirectly determines the number of clusters.

V COMPARATIVE ANALYSIS OF VARIOUS ALGORITHMS

For performance evaluation of the most popular clustering techniques K-Mean clustering, Fuzzy C-Means clustering, Competitive Neural Network and Quantum 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.

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.

The *leukaemia data set* contains expression levels of 7129 genes taken over 72 samples. Labels indicate which of two variants of leukaemia is present in the sample (AML, 25 samples, or ALL, 47 samples).

VI EXPERIMENTAL SIMULATION AND RESULTS

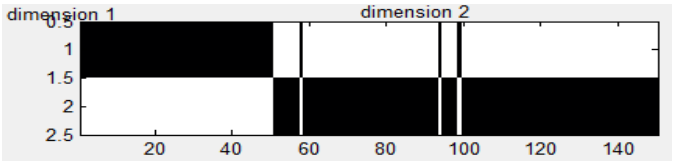
We applied several of the most commonly used clustering algorithms for gene expression data. By analyzing the results of this software, we observe significant variations in performance. In the following we compare the performance on different datasets.

TABLE I
TESTS ON IRIS PLANTS DATASET

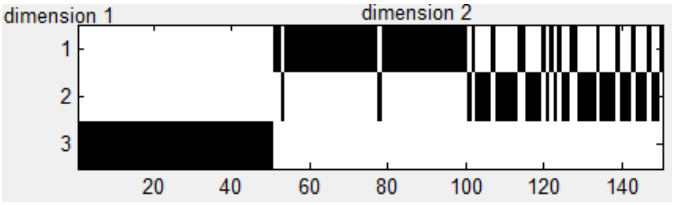
Method	2 Clusters	3 Clusters
K- Means	J = 0.53996	J = 0.34714
Fuzzy C-Means(FCM)	J = 0.53996	J = 0.33781
Competitive Net	J = 0.51678	J = 0.33611

The Jaccard score (J) reflects the 'intersection over union' between the algorithm and 'real' clustering, and its values range from 0 (void match) to 1 (perfect match).

As we know that from Silhouette plot analysis, for fisher iris data the best value of K is 2 which provides the best clustering so it is shown in TABLE I with the help of Jaccard Scores the comparative analysis of various clustering techniques depicting the best value of K and the clustering technique best suited to it.



(a)



(b)

Fig 3 A Classification Alternative Display of K-Means Clustering Technique with two different values of K. Here location on the x-axis equals its location in the Real classification and in the Data matrices, and its location on the y-axis represents the tag proposed by the algorithm. Perfect classification is therefore represented as homogenous rectangles. (a) K-Means with K=2 (b) K-Means with K=3.

TABLE II
TESTS ON IRIS PLANTS DATASET COMPARING WITH QUANTUM CLUSTERING

Preprocessed Data

Method	Jaccard Scores
K-Means(4 Clusters)	0.25378
Fuzzy C-Means(FCM) (4 Clusters)	0.25378
Competitive Net (4 Clusters)	0.25539
Quantum Clustering(QC) (σ = 0.599)	0.35329

TABLE III
TESTS ON LEUKAMIA DATASET SHOWING THE EFFECT OF APPLYING PREPROCESSING

METHOD	JACCARD SCORE
Raw Data (for 2 clusters)	
K-Means	0.49296
Fuzzy C-Means(FCM)	0.49335
Competitive Net(CNN)	0.50274
Quantum Clustering(QC)	NA
Preprocessing(SVD) (truncation to 5 dimensions)	
K-Means	0.63419
Fuzzy C-Means(FCM)	0.51213
Competitive Net(CNN)	0.64945
Quantum Clustering(QC)	0.93395

TABLE III shows that applying the selected algorithms using raw data (i.e. without SVD pre-processing) yields poor outcomes. Next we applied the SVD pre-processing step selecting and normalizing the 5 leading SVD components ('eigengenes' according to Alter, [10]) thus reducing the

matrix from 7129X72 to 5X72. Clustering has improved after dimensional truncation, yet not all algorithms correctly cluster the samples. Note that only QC shows a substantial degree of consistency with the 'real' classification.

VII CONCLUSIONS

In this paper we demonstrate how different clustering algorithms may lead to different results. The advantage of COMPACT is in allowing many algorithms to be viewed and evaluated in parallel on a common test set. Through COMPACT one can evaluate the impact of changing the algorithm or its parameters (e.g., sigma value in QC, number of iterations for the Competitive Neural Network, starting points of K-Means, Fuzzy C-Means and more). Being able to run a number of clustering algorithms, observe their results (quantitatively and graphically) and compare between them is beneficial for researchers. We find it advisable to start with a problem that has a known classification (referred to as 'real classification') and use the statistical criteria (i.e. efficiency, purity and Jaccard score) to decide on the favourable clustering algorithm. For general research problems, where no known classification exists, the same statistical tools may be used to compare results of different clustering methods with one another. We presented here a comparative analysis of some well-known clustering methods with one relatively new method, QC.

The advantages of this software are:
(i) presenting an integrative, light package for clustering and visualization,
(ii) integrating an efficient compression method and
(iii) introducing the QC algorithm as part of the available clustering options.
This software is very useful for data analysis.

ACKNOWLEDGEMENT

I am thankful to Dr. Kamal Mehta(Head of Department), Computer Science & Engineering for giving thoughtful suggestions during my work. I owe the greatest debt and special respectful thanks to Dr. P. B. Deshmukh (Director), Shri Shankaracharya Technical Campus, Bhilai for their inspiration and constant encouragement that enabled me to present my work in this form.

REFERENCES

[1] Z. HUANG, ―*Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values ",* Kluwer Academic Publishers, *1998*.

[2] A. Mustafa, A. Akbar, and A. Sultan , ‖ *Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization",* International Journal of Multimedia and Ubiquitous Engineering , Vol. 4, No. 2, April, 2009 .

[3] Li Wei, Nitin Kumar, V. Lolla and H. V. 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] F. Altiparmak, H. Ferhatosmanoglu, S. 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.

[5] E. Hajizadeh, H . Davari Ardakani and J. Shahrabi:"*Application of data mining techniques in stock market*", Journal of Economics and International Finance Vol. 2(7), pp. 109-118, July 2010.

[6] K.-L. Du "*Clustering : A Neural Network Approach* ", www.elsevier.com/locate/neunet.

[7] D. HORN AND I. AXEL, "*Novel clustering algorithm for microarray expression data in a truncated svd space*", Bioinformatics, 19 (2003), pp. 1110–5. Journal Article England.

[8] Ren Jingbiao, Yin Shaohong, "*Research and Improvement of Clustering Algorithm in Data Mining*" 2nd International Conference on Signal Processing Systems (ICSPS), 2010.

[9] Al-Zoubi, M.B., A. Hudaib and B. Al-Shboul," *A fast fuzzy clustering algorithm*". Proceedings of the 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, February 2007, pp. 28-32.

[10] Alter, O., Brown P. O, Botstein D.," *Singular value decomposition for genome-wide expression data processing and modelling "* Proc Natl Acad Sci U S A. 2000, 97: 10101-10106.

