International Journal of Applied Research 2017; 3(12): 521-524

ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 5.2
IJAR 2017; 3(12): 521-524
www.allresearchjournal.com
Received: 20-10-2017
Accepted: 21-11-2017

Geetanjali
M. Tech Student (CSE)
PPIMT College, GJU University, Hisar, Haryana, India

Neeraj Verma
Assistant Professor, CSE Department, PPIMT, Hisar, Haryana, India

Paru Raj
HOD, CSE, Department, Ppimt, Hisar, Haryana India

Correspondence
Geetanjali
M. Tech Student (CSE)
PPIMT College, GJU University, Hisar, Haryana, India

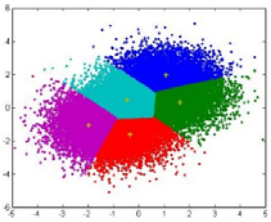
Mining patterns for clustering using modified k-means and SVM (Support Vector Machine)

Geetanjali, Neeraj Verma and Paru Raj

Abstract
Data mining can be termed as a process of extracting patterns (knowledge) and modelling query from data which is stored in database. Classification is one among of its concept and techniques. K-means is very popular clustering algorithm used. And for classification there are many algorithms provided.
This research article is proposing a novel hybrid mining approach by using modified K-Means and Support vector machine algorithm. Modified K-Means utilized here for making the clusters from given dataset and SVM is utilized for classification (on clustered dataset obtained from modified K-means clustering). Experiments are performed over different datasets which are taken from UCI Irvin. Datasets which are used for comparing clustering algorithm are provided in Table 1 along with their details. Evaluations are done on different datasets of following parameters: Accuracy obtained from new algorithm using confusing matrix which is being created for every dataset and time taken to execute the task. Experimental results show that the proposed algorithm has better accuracy than the traditional algorithms. Results gives better and modified classification of clustered data. Additionally, proposed algorithms provide better clustering result than the existing algorithms.

Keywords: modified k-means, clustering, support vector machine (SVM), classification, confusion matrix

1. Introduction
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, but are very dissimilar to objects in other clusters. A main problem that frequently arises in a great variety of fields such as data mining and knowledge can discovery, with data compression and vector and pattern recognition with pattern classification is the term of clustering problem. It too has been applied in a large variety of applications, for example, image segmentation, object and character recognition, There are more approaches in including splitting and merging process and randomized approaches, all methods based on symmetry process.
One of the most popular and widely studied clustering methods that minimize the clustering error for points in Euclidean space is called K-means clustering. K Mean classify a given data set through certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. It is well known that the basic K-means algorithm does not produce an analytic solution. The iterative process is only guaranteed to converge to a local rather than a global solution. The solution will depend on how the objects are initially assigned to clusters; this aspect has already been explored by various authors. The K-means algorithm gave better results only when the initial partition was close to the final solution. Several attempts have been reported to solve the cluster initialization problem.



Data mining can be defined as a process of modelling query, obtaining patterns and information from data. Data mining has many technologies including machine learning and parallel processing.

Classification is one of its concepts and techniques. Classification can be defined as the data mining function in which records can be grouped into some significant subclasses. Aim of the classification is to forecast the class for data. Labels of categorical class can be predicted by classification.

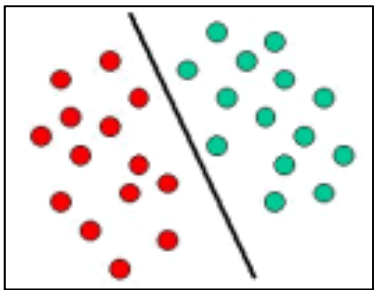


Fig: classifications

2. Related Work

K-means algorithm

K-means is frequently used algorithm for clustering. By using any of the criterion Inter-clustering or intra-clustering K-means can be measured. O (PCI) is the time complexity for K-means algorithm, where P-is the number of input patterns and C-is the he numbers of clusters needed and I-is the number of iterations to complete clustering.

1. Take a set S of N patterns $(x_1, x_2 \ldots x_n)$ having their dimension's s is divided into Tclusters which can be denoted by$c_{1,} c_2 \ldots c_T$.
2. *M* is the cluster metrics which becomes minimum. And the equation is represented as:

$$M(c_1, c_2 \ldots c_T) = \sum_{T=1}^{T} \sum_{x_{j \in c_T}} \| x_j - c_T \|^2$$

3. Then assign data to representative center upon minimum distance.
4. Start computing the new cluster center.
5. Process will terminate when cluster center become stable for two consecutive iterations.

*Clusters are obtained in result in Figure 2.

This is the standard algorithm used for k-means. Where, Number of clusters are taken as T. Cluster metrics is M. Equation is used to measure the minimum distance between the data points.

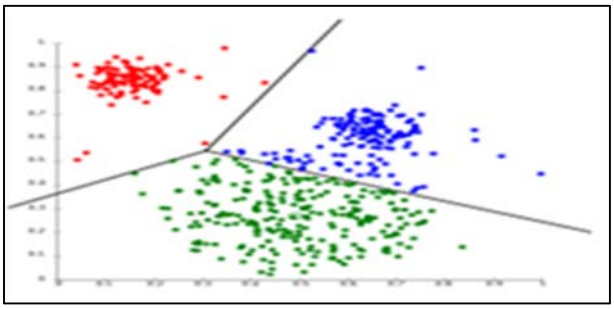


Fig. K-means clustering

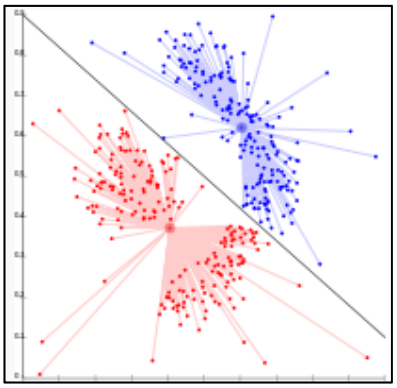


Fig: k-means for density based clustering

Problem Statement

Data mining is a process in which ant type of query can be modelled and patterns can be retrieved using proper technique on the dataset. As data mining is a big area of research so, single algorithm cannot be applied for the various problems in data mining. I proposed a hybrid approach which is applied to classify the data. Analysis of large dimensions of data is also done in the paper. Two techniques are combined in the paper to classify the data. Comparison of the existing algorithm and proposed algorithm gives the conclusion. These two techniques are: Modified k-means algorithm and SVM. For clustering k-means is used. And SVM is used here for classification. Initially, dataset is loaded and then clustered using modified k-means algorithm. SVM algorithm is used in the paper for classifying the data which is clustered by k-means. It divides the data level wise. Results are now obtained from the experiment. There were two attributes selected for the comparison. They are the accuracy and time taken to complete the task on the number of instances taken in the system. Two separate tables were created to store the results. Now, this modified k-means is compared with the original algorithm. It was found from the observation that the modified is better than the traditional algorithm. Because the time it takes to perform the complete task is less than the time taken by the traditional. This is now applied to different dataset to check the validity of the proposed system. Results are again compared with the traditional algorithm. And finally, it is proved that the proposed algorithm has better accuracy and it take less time to execute any operation.

3. Proposed Protocol

Proposed System

Many algorithms have been introduced which create numerical patterns and these patterns are extracted from collection of unsupervised decision tree.

Instead of using different approach and algorithms for every task my proposal is to combine two algorithms and they are Modified K-Means for clustering that concentrates just the subset of valuable features for clustering and SVM for classification. It is a novel hybrid approach using modified k-means and SVM which support vector machines. Proposed system work on every data of numerical dataset. Numerical data is preprocess and then it is clustered using modified k-means.

There are many clustering techniques available for clustering, but this system will work on modified k-means only. Because in k-means we choose random clusters at the

beginning. But this does not happen in modified k-means. In this technique, selection of clusters is not random. The distance between two data points is measured and the datasets having minimum distance to each resides in one cluster.

Among classification techniques available SVM is used for the classification. Because when compare to other classification technique SVM gives better accuracy.

Confusion matrix's concept is also used with this technique to measure the accuracy of the datasets.

Flowchart for the Proposed System

Given below in Figure is the flowchart for proposed process

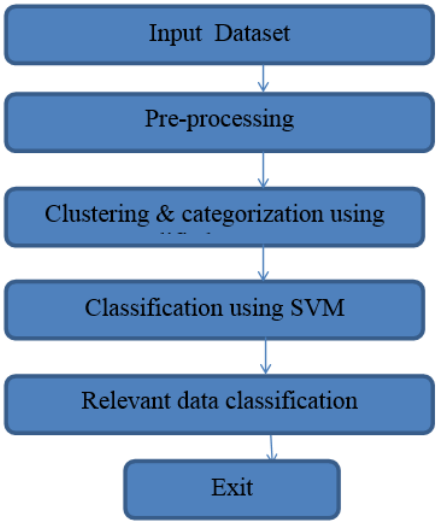


Fig: Flowchart of proposed system

Initially, Numerical Datasets are loaded from UCI repository. Datasets are clustered using Modified K-means algorithm and hence retrieve different clusters for different datasets. Then the data is classified on clustered datasets using SVM.A confusion matrix is created to understand the results clearly.

4. Results and Discussions

This research selects Seven datasets and they are:

1. Clouds
2. Magic
3. Knowledge
4. Seeds
5. Glass identification
6. ILPD
7. Iris.

They all are saved in seven different text files. Initially, a SRS.m file is created, which have all the links available related to perform this hybrid approach of clustering and classification. Once SRS is run, all other classes will be automatically linked to this class. When we run the SRS.m file, it gives SRS figure in the output. This figure will appear on the screen which has three push buttons: loading dataset, clustering and classification and accuracy space edit bar. When loading dataset is clicked, it will load the dataset which is selected for performing the experiment. The loaded dataset will appear on the SRS figure, we created for execution. For example, if iris dataset is selected for loading the dataset then the data of iris will be loaded along with its attributes. Iris has four attributes and they are: petal length, petal width, sepal length, sepal width. Now, next button which is clustering is clicked to perform the clustering on the selected dataset. After performing the clustering, one more figure is displayed on the editor. This figure will display the clustering results for the selected dataset. After this classification is performed on clustered data, by clicking the push button for classification. After performing the classification, it will display a SRS figure which returns a confusion matrix and gives the accuracy results in the percentage.

Results for Iris Dataset

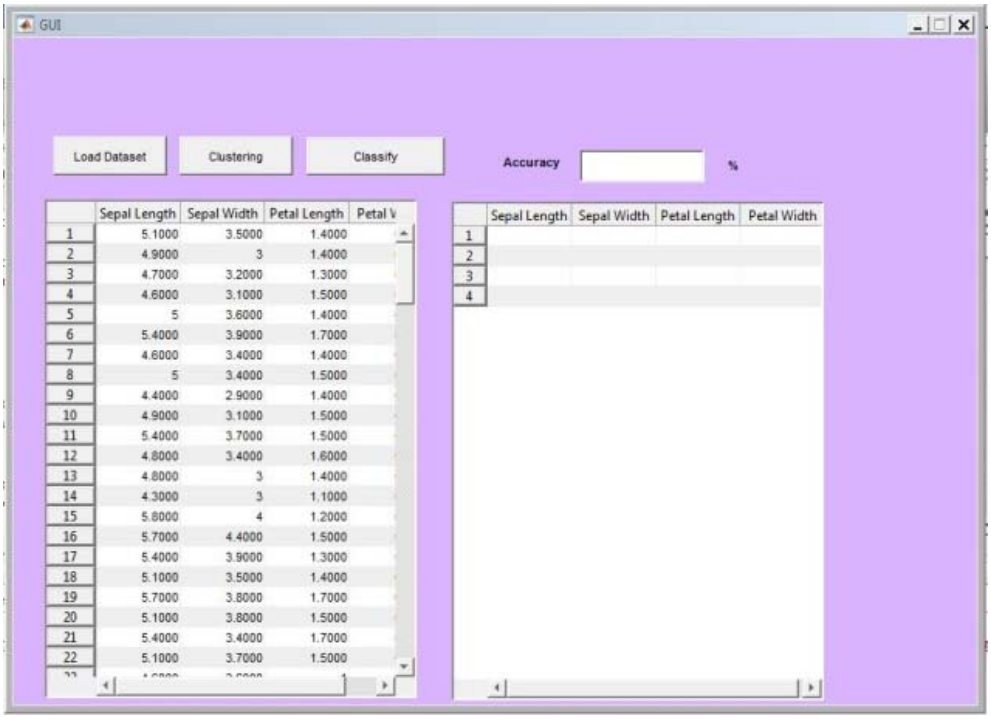
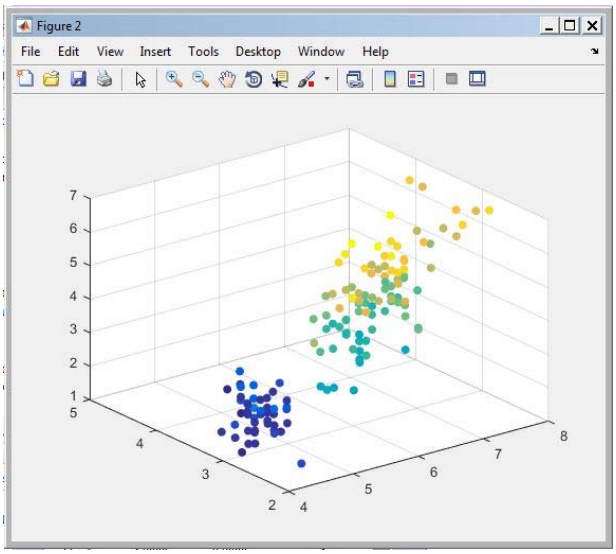


Figure is displayed in the output when GUI. m class is being run for iris dataset. Iris has four attributes and they are: petal length, petal width, sepal length, sepal width. Now, next button which is clustering is clicked to perform the clustering on the selected dataset. Clustering is then performed using Modified K-means algorithm. Figure given below demonstrates the clustering performed on iris dataset.



5. Conclusion

Clustering is a typical technique for arithmetical research of information, which is used by various fields of life, as bioinformatics, user reviews, picture examination, design acknowledgment and machine learning. We built up a novel hybrid approach for data mining which depends on combination of two algorithms and they are the modified k-Means and SVM (Support Vector Machine). The modified k-Means method is first divides the dataset into different clusters and for classification SVM is used. Proposed algorithm obtains better clustering results than existing ones. And, the proposed hybrid approach gives better clustering results then other clustering algorithm.

6. References

1. Gutierrez-Rodriguez AE, Fco Martinez J. Trinidad, M. Garcia-Borroto, J.A. Carrasco-Oacha," Mining patterns for clustering on numerical datasets using unsupervised decision tree, Knowledge based systems. 2015; 82: 70-79.
2. Michalski RS, Stepp RE. Automated constructions of classifications; conceptual clustering versus numerical taxonomy, IEEE Trans. Pattern Anal. Machine Learn. 1983; 5(4):396-410.
3. Bing Liu, Yiyuan Xia, Philip S. Yu, Clustering through Decision Tree Construction, (CIKM-2000), Washington DC, USA. 2000, 6-11.
4. Daxin jiang, Jai pie (CANADA), Aidong Zhang(USA), "General approach to mining quality based clustering on microarray data, L. Zhou, B.C. Ooi, and X. Meng (Eds.): DASFAA, LNCS 3453, Springer-Verlag Berlin Heidelberg, 2005, 188-200.
5. Vivekanathan P. Different data mining algorithm: A Performance Analysis/ 2012; 1(3):79-84.
6. Muhammad Ali Masood, Khan MNA. Clustering Techniques in Bioinformatics", I.J. Modern Education and Computer Science. 2015; 1:38-46.
7. Hanumantha Rao K, Srinivas G, Ankam Damodhar, Vikas Krishna M. Implementation of Anomaly Detection Technique Using Machine Learning Algorithms, International Journal of Computer Science and Telecommunications. 2011; 2(3):25-31.
8. Sumit Garg, Arvind Sharma K. Near Analysis of Data Mining Techniques on Educational Dataset, International Journal of Computer Applications. 2013; (0975-8887):74-5.
9. Lior Rokach, Oded Maimon (Eds.), Clustering Methods, Data Mining and Knowledge Discovery Handbook, XXXVI, 1383. illus., Hardcover ISBN:0-387-24435-2, 2005; 400.
10. Pavel Berkhin. Survey of Clustering Data Mining Techniques, Grouping multidimensional data-recent advances in clustering, ISBN 9873-3-540-28348-5. 2006, 25-71.
11. Muhammet Mustafa Ozdal, Cevdet Aykanat. Hypergraph Models and Algorithms for Data-Pattern-Based Clustering, Data Mining and Knowledge Discovery, Kluwer Academic Publishers. Manufactured in The Netherlands. 2004; 9:29-57.
12. Mythilli, Madhiya. An Analysis on Clustering Algorithms in Data Mining. CSMC. 2014; 3(1):334-340.
13. Arthur Zimek, Ira Assent, Jilles Vreeken. Frequent Pattern Mining Algorithms for Data Clustering, DOI 10.1007/978-3-319-07821-2_16. Springer International Publishing Switzerland. 2014, 403-423.
14. Tamizharasi K, Dr. Uma Rani, Rajasekaran K. Performance analysis of various data mining algorithms. International Journal of Computing Communication and Information System (IJCCIS). 2014; 6(3):118-127.
15. Jiangping Chen, Ting Hu, Pengling Zhang, Wenzhong Shi. Trajectory clustering for people's movement pattern based on crowd souring data, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, ISPRS Technical Commission II Symposium, Toronto, Canada. 2014, 6-8.
16. Keerthana P, Thamilselvan P, Sathiaseelan JGR. Performance Analysis of Data Mining Algorithms for Medical Image Classification. International Journal of Computer Science and Mobile Computing. 201; 5(3):604-609.