International Journal For Technological Research In Engineering
Volume 4, Issue 11, July-2017 *ISSN (Online): 2347 - 4718*

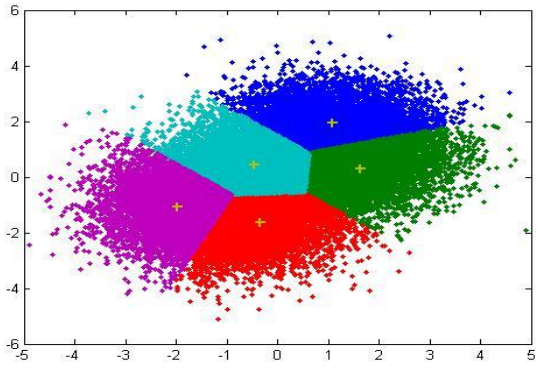
DATA MINING ALGORITHM USING K-MEANS CLUSTERING TECHNIQUES

Indu[1], Anjali Namdev[2]
[1]PG Student, [2]Assistant Professor, CSE Department, S(PG)ITM Rewari, Haryana, India

I. INTRODUCTION

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.

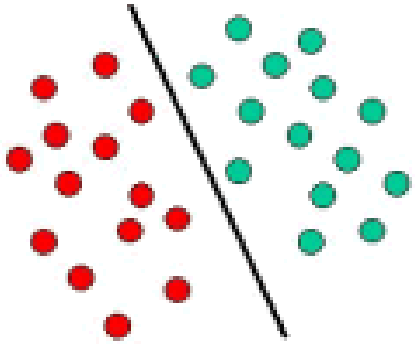


Fig : classification

Figure is used here to display how the basic classifier separate the two classes.

Clustering is the technique in which set of objects gets divided into different clusters and these clusters should not be overlapped with each other. The main motive of clustering should be that the one cluster has similar type of objects than the other cluster. A set of patterns are obtained in results of pattern-based clustering that describes each cluster and gives a description of objects belonging to each cluster. An expression that describes an object subset is called a "Pattern". There is no specific algorithm or technique available which can be applied to obtain better efficiency. Different algorithms are grouped to resolve any problem.

Among most recent algorithms of data mining, some are applicable to the numerical dataset. These datasets need to discretize before applying any algorithm on it. And It also may happen that the important information of the dataset will be lost, if we apply discretization on dataset before applying any algorithm on it. Data processing is important when the available data is incomplete and dirty. There may be noise in data or it also may happen that the available data is inconsistent. Many data processing techniques are available to make the data clean. It can be achieved by data integration, data cleaning, by applying transformation on data, data reduction and by discretization. Now discretization also has many methods to choose from. Binning is the simple discretized method to make the data clean. Equal width portioning is used to divide the available data into specific number of gaps. All these gaps are of equal size. These intervals are uniformly distributed.Width of this gaps are determined by the lowest and the highest value of the attributes. In equal-depth partitioning same number of samples gets distributed from a available range of data.

Discretization can handle three types of data. They can be: any numeric value whose set is unordered, any ordinal value whose set is ordered and it can be any continuous value having real numbers. By this the size of data can be reduced and modified data can be used for the future evaluations and analysis. Another algorithm is also proposed which do not need a discretization. Unsupervised patterns are used to create the pattern but clustering is done by using traditional algorithm" K-Means". K-Means sometime may create empty clusters which may lead to the anomalous behavior of the system. To overcome this problem, I proposed a hybrid approach with "Modified K-Means algorithm" for clustering which assures the creation of clusters which are never empty and for classification SVM (Support Vector Machine) are used.

II. 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 SVMfor 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[21]. 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 Fig 5 is the flowchart for proposed process:

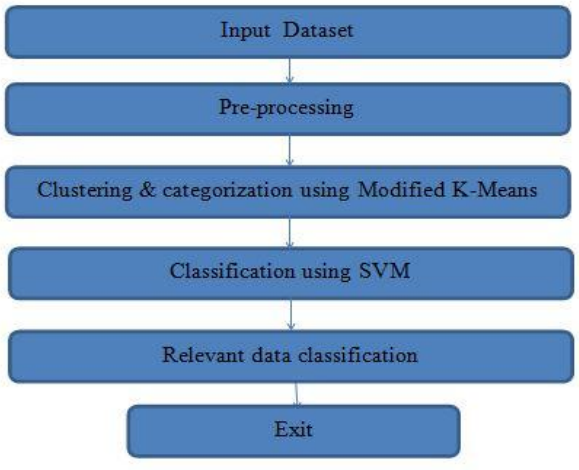


Fig. Flowchart of proposed system

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.

Proposed Algorithm Details:

Modified approach K-mean algorithm:

K-mean algorithm is widely used clustering algorithms. This algorithm gives effective results when used with small datasets, but cannot handle the largedatasets. Another problem with this algorithm is that it produces the empty clusters due to the selection of center vector. This leads to anomalous behavior of system and significant performance degradation. To address this problem, the modified algorithm is proposed which assures the creation of non-empty clusters. In modified K-means, the computation of new center vector is different from k-means. Old center vectors areassumed that they are already the member of the concerned clusters. Data items are already allocated to their specific clusters.

Algorithm:

Modified (M,q), M = {x1, x2 … xn}

The number of clusters q1 (q1<q) and a dataset containing o number of objects(xijk)

Output: A set of C clusters (Cij) that minimize the Cluster - error criterion.

Algorithm:

1.Initially, distance between each data- points is measured and store it in the set T.

2.A data-point set Dp ($1 \leq l \leq q+1$) is formed after finding the closest data points from T.

3.Delete these two data points from the set we make (T).

4. Closest data point to data point set Dt is selected from T and added to Dt. Also, delete it from T.

5. Repeat step 4 until the number of data points in Dp reaches($n \backslash q$).

6. If $l < q+1$, then$l = l+1$,select another pair of dataset from T which has the shortest from the data point of Dt distance Go to step 4.

Support vector machine:

SVM's were originally invented by Vapnik&Chervonenkis in 1963. Support vector machines are an arrangement of managed learning strategies utilized for order anomalies discovery. SVM is supervised learning model.

III. TOOLS AND TECHNOLOGY

HARDWARE REQUIREMENTS:

Operating system	64 bit
Processor	PC with anIntel-R Pentium II-class processor, 2.16 Hz
RAM	256 MB (Recommended)
Hard disk	10GB (8GB space required to install MATLAB and related software's)

SOFTWARE REQUIREMENTS

For Implementation:

MATLAB (8.2.0.701 (R2013b)), developed by Math works.

System Software

The software can be executed in the Windows 2000 (Professional or Server) or Windows XP Professional environment.

MATLAB was developed by Math Works. MATLAB stands

for MAT-matrix LAB-laboratory. MATLAB is platform independent and hence can be operated on any operating system.
Initially, it was started with very normal matrix manipulation and now it has grown so much. It provides online help facility to make the implementation easy.
MATLB have great capabilities and they are:

- It can handle numerical computations.
- It has good data access.
- It provides data analysis and visualization.
- It can be used for program and algorithm development.
- Applications can also by developed and deployed.

Hardware's like TMS320C6713, NIDAQ, Arduino and FPGA are some which can be interfaced with MATLAB and data capabilities can be increased using them. IT has better hardware connectivity such that it can be connected to any standalone hardware systems.
MATLAB also has very better data access that it can access any type of data like database, excel, standard data and scientific data in whatever form it is. Data can be analyzed and visualized in MATLAB.It also provides toolbox for to solve any problem.
MATLAB has predefined functions in its library. It provides its own library and this can help in efficient any easy programming.
Application of MATLAB:

- Can be used in technical computing.
- Used in control design of application and programs.
- Used for communication and DSP design.
- Can be used for test and measurements.
- It is used for image processing.
- Also, used in finance analysis and modeling.

MATLAB also has some pro's and con's, and they are:

- MATLAB is very expensive.

Solution to the problem: Use Student version of MATLAB.

- It has slow speed.

Solution to the problem: By doing proper structuring.

Performance:
It is special technical programming language, using this language one can perform:

- Scientific calculations
- Engineering calculations

MATLAB is better than other languages like C, FORTRAN. Programs are written in this can be easily run and complied as it has its own complier. Work space is temporary memory. Edit / debug window is used to create or edit M files. All files in this language has ".m" extension.

IMPLEMETATION:
Coding for the proposed system has been done in MATLAB. Steps involved in executing the experiment:

- Dataset is loaded.
- Data is pre-processed for clustering.
- Clustering is then performed on the dataset using modified k-means algorithm.
- Classification using SVM.
- Relevant data is classified.
- Terminate the program.

Experiment is performed on datasets listed as follow: Clouds, Glass-identification, LIPD, Iris, Knowledge, Magic, Seeds, Sonar and wine which re taken from UCI repository. Details of datasets which are used for comparing is provided in table 1.
Implementation is done in MATLAB (8.2.0.701 (R2013b), developed by Math works. This is used in various field and we are using it for Program and Algorithm development as it adaptable for data in any form. This experiment was executed on a desktop PC and 64-bit operating system and Intel-R Pentium Processor at 2.16 Ghz.
To execute the coding ten files are created and they are listed as follow:

1. ttest_selection.m

ttest_selection.m

Test selection class has a function(features, labels, B).

2.Train_MSVM.m

Train_MSVM.m

Train_MSVM class has function (Train data, Train label).

3.irisdataset.txt

irisdataset.txt

This text folder has data of iris. Like this every dataset has its own details and folders. There are seven different datasetname.txt files in this program.

4. SRS.m

SRS.m

It is graphic user interface (SRS) file. It has one output function and one opening function. It has three push buttons as push button_1, push button_2 and push button_3. It has two edit as edit1_call back and edit1_Create function. All other files are linked with this file.

5. SRS.fig

It opens the SRS screen by which user can interact with the figure. A SRS figure is displayed on the screen having different buttons of loading datasets, clustering, classification, accuracy percentage. Once the SRS.m file is run, it will open this file automatically. Now, dataset having different features and objects will be loaded after clicking the button. Then, it is clustered, when clustering is clicked. A MATLAB figure is displayed on the screen which displays the clustering results. After that it will classify the data and will give the accuracy percentage.

6. F_SVM.m

F_SVM.m functionhas F_SVM (data, label, sigma,nb_cvfolds).

7.crossval_svm.m

This class is created for cross validations performed on datasets. This function is created to train an SVM model overtraining instances. Function is: crossval_svm(data, label, sigma, k).

8. Confusion_matrix.m

Confusion matrix function is: confusion_matrix (predicted, labels, classes names).Confusion matrix for all datasets are created separately. Confusion matrix is the method for evaluating the learning algorithms. Confusion matrix has two classes: one is predicted class and second is actual class. In a confusion matrix, there are two types of instances: correctly classified instances and incorrectly classified instances. Confusion matrix has two performance evaluation class label: Yes and No. for every class label there is four functions and they are: precision, Recall, False positive rate and F-measure.

9. Classify MSVM.m:

This is the classifier function. Function is: Classify_ MSVM (test_data, label, svmstruct, level). Clustered data is classified using this function. SVM is used for classification for the proposed system.

10. Best_features.m

Best feature function is: Best_Feature (mat, label, sigma).

Like this, classes for all the datasets are created. All seven datasets (Clouds, magic, knowledge, seeds, ILPD, glass identification and iris) are saved in seven different text files. A SRS.m file is created at the starting 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 DTASETS
- CLUSTERING
- CLASSIFICATION

When loading dataset is clicked, it will load the dataset which is selected for performing the experiment. One text file is selected from seven available datasets. The data set of the same is loaded to the command window of the MATLAB. 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. As MATLAB provides tools and technologies to display the bars, graphs and other analysis. 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. Using the same process, all datasets are executed one by one and the accuracy results are obtained. These results are now compared with the results of previous researches. Comparison results will conclude the motive of the research.

IV. RESULT AND DISCUSSION

This research selects Seven datasets and they are:

- Clouds
- Magic
- Knowledge
- Seeds
- Glass identification
- ILPD
- 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:

IRIS FILTER RESULT					
	Sepal length	Sepal width	Petal length	Petal width	Species
14	4.3	2.0	1.3	0.3	Setosa
9	4.4	2.8	1.4	0.1	Setosa
39	4.4	2.9	1.4	0.1	Setosa
42	4.4	3.0	1.3	0.2	Setosa
41	4.5	2.9	1.5	0.2	Setosa
5	4.6	3.1	1.4	0.2	Setosa
6	4.6	3.5	1.6	0.1	Setosa
47	4.6	3.6	1.4	0.2	Setosa
4	4.7	3.2	1.5	0.1	Setosa
9 result note out of 150					

Fig. loading iris dataset

Figure 6 is displayed in the output when SRS.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 7 given below demonstrates the clustering performed on iris dataset.

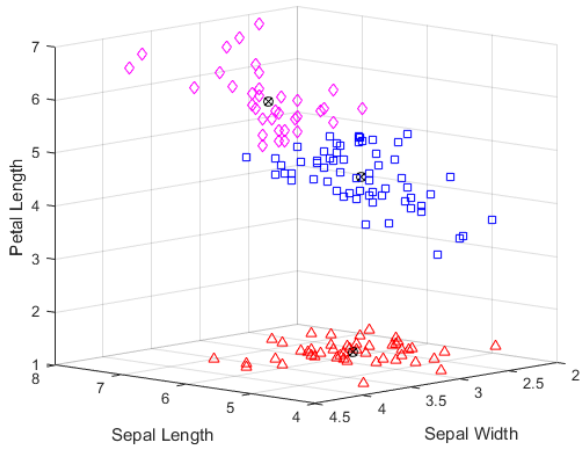


Fig. clustering using k-means

After clustering, SVM is used clustered patterns for classification which is shown in figure . Figur displays the accuracy results which get added to the edit bar of the figure. It will display the accuracy in percentage along with its all attributes we taken for clustering. Accuracy is noted down because this system will compare the accuracy of current research to the previous research.

IRIS FILTER RESULT					
	Sepal length	Sepal width	Petal length	Petal width	Species
14	4.4	2.1	1.4	0.1	Setosa
9	4.45	2.4	1.4	0.3	Setosa
39	4.4	2.2	1.4	0.1	Setosa
42	5.7	3.0	1.3	0.3	Setosa
41	4.8	2.9	1.2	0.2	Setosa
5	4.7	3.1	1.4	0.2	Setosa
6	4.	3.6	1.5	0.2	Setosa
47	4.6	3.6	1.4	0.2	Setosa
4	4.7	3.1	1.1	0.3	Setosa
Modified iris result					

Fig. loading iris dataset

Figure also shows the classification result. It displays a

confusion matrix created for iris data after classification. Confusion matrix will be displayed in another SRS figure whose specifications are provided in coding. Confusion matrix are frequently used methods when anytraining set is classified. ROC curve is also used for the purpose. But for this research confusion matrix are selected because it provides more accurate results then the ROC curve.
Confusion matrix has two classes: predict class and actual class.
To find the accuracy it uses a formula, which is: Accuracy = {(TP + TN)/ (TP+TN+FP+FN)}.
Where,TP is True Positive.
TN is TRUE Negative.
FP is False Positive
And, FN is False Negative.

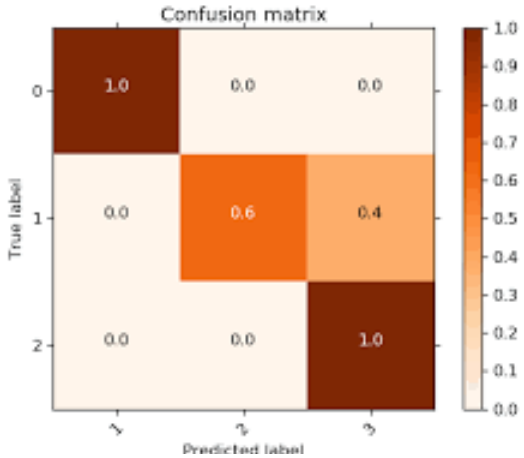


Fig. Confusion Matrix of iris dataset

This is the confusion matrix created for iris dataset. Accuracy results obtained are then stored in the table 2 which is created to store accuracy percentage for every dataset. These results are now compared with the results of previous researches. Comparison results will decide that the hybrid approach of clustering and classification is better than the traditional or not.

RESULTS FOR GLASS IDENTIFICATION DATASET:
Figure 10 shows the classification which is being performed on another dataset which is glass identification dataset.

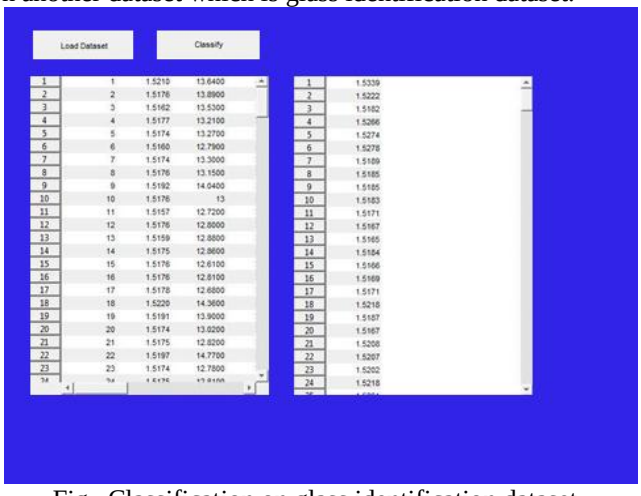


Fig. Classification on glass identification dataset.

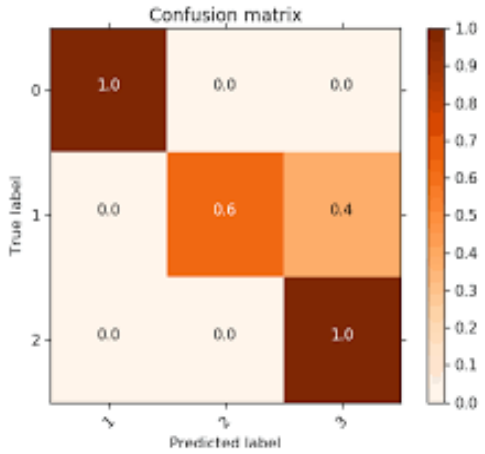


Fig. Confusion matrix of glass identification dataset

Same as these datasets, all datasets are executed one by one.
ACCURACY RESULTS:
A table 2 is hence created after performing the experiment and values obtained are stored to the table. Accuracy obtain from every dataset is stored for every dataset. They are stored against their name in table 2. Table 2 also contains the objects on which this experiment is performed which is a novel hybrid approach of k-means clustering and SVM for classification.

TABLE 2
ACCURACY OF DATA SET

DATASET	OBJECTS	ACCURACY
CLOUDS	101	84.2
GLASS IDENTIFICATION	192	89.3
ILPD	601	87.88
IRIS	152	94.22
MAGIC	192	92.30
SEEDS	223	89.22
KNOWLEDGE	427	87.3

V. 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 algorithms.

