Volume 3, Issue 9, September 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Comparative Study on Feature Selection Using Data Mining Tools

Ritu Ganda*
Department of Computer Science
J.C.D.M college of Engineering and Technology
Guru Jambheswar University of Science and Technology
India

Vijay Chahar
Asst Professor, Department of Computer Science
J.C.D.M College of Engineering and Technology
Guru Jambheswar University of Science and Technology
India

Abstract— Clustering is a important technique of data mining. Clustering is an unsupervised learning problem that group objects based upon distance or similarity. Each group is known as a cluster. In this paper we make use of a large database 'Cardiology Dataset' containing 14 attributes and 303 instances to perform Feature Selection on K-means algorithm. We compared the results of simple clustering technique and clustering (K-means) with feature selection for Cardiology dataset, based upon various parameter using WEKA (Waikato Environment for Knowledge Analysis) and TANAGRA data mining tools. The results of the experiment show that clustering with feature selection give promising results on WEKA with utmost accuracy rate and robustness.

Keywords— Data Mining, K-means, Cfs filtering, WEKA, TANAGRA, Cardiology Dataset.

I. INTRODUCTION

Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information [2]. The main aim of feature selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. In many real world problems Feature selection is must due to the abundance of noisy, irrelevant or misleading features. For instance, by removing these factors, learning from data techniques can benefit. Features can be characterized as:

1. **Relevant**: These are features which have an influence on the output and their role cannot be assumed by the rest.
2. **Irrelevant:** Irrelevant features are defined as those features not having any influence on the output, and whose values are generated at random [4].

A typical feature selection process consists of four basic steps (shown in Fig.1), namely, subset generation, subset evaluation, stopping criterion, and result validation [3]. Subset generation is a search procedure [5][3] that produces candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation criterion. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied. Feature selection can be found in many areas of data mining such as classification, clustering, association rules, and regression. Feature selection algorithms designed with different evaluation criteria broadly fall into three categories: the filter model [7], the wrapper model [8], and the hybrid model [9]. The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the mining algorithm aiming to improve mining performance, but it also tends to be more computationally expensive than the filter model [6][7]. The hybrid model attempts to take advantage of the two models by exploiting their different evaluation criteria in different search stages.

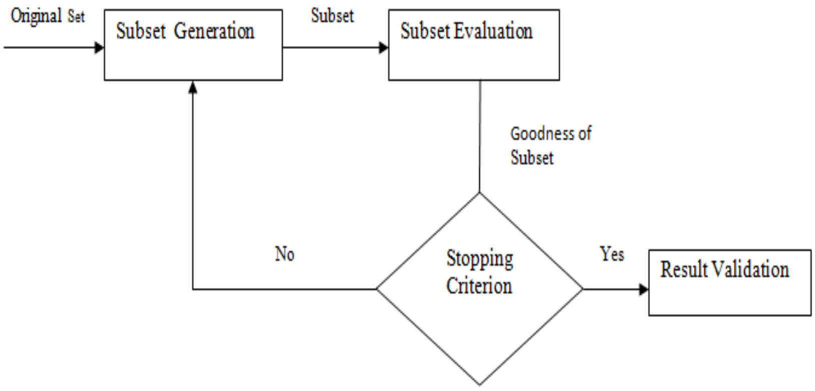


Fig 1 Four Key Steps of Feature Selection

A. Organisation of the paper

The paper is organized as follows: Section 2 describes K-means algorithm. Section 3 defines problem statement. Section 4 describes the proposed clustering method to group the object based on distance or similarity measure and clustering with feature selection technique of data mining. Experimental results and performance evaluation are presented in Section 5 and finally, Section 6 concludes the paper and points out some potential future work.

II. K-MEANS ALGORITHM

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The K-means algorithm is a classical clustering method which is used to group large datasets into clusters [1]. It is the unsupervised classification to find optimal clusters. The algorithm is often considered to be a partitioning clustering method, and it works as follows. It arbitrarily chooses the cluster centre then the objects are assigned to the similar cluster, which are more similar. The cluster means are updated for each cluster until there is no change. The disadvantage of using K-means method is the number of cluster should be specified in the beginning and it is not able to generate the cluster with different shapes.

Pseudo code for K-means algorithm is:

```
1. Select k points as the initial centroids.
2. repeat
3. Form k clusters by assigning all points to closest centroids.
4. Recompute the centroid of each cluster.
5. Until the centroid don't change.
```

III. PROBLEM STATEMENT

The problem in particular is a comparative study of clustering technique algorithm K-Means with and without feature selection on Weka and Tanagra using Cardiology Dataset consisting of 303 instances and 14 attributes.

IV. PROPOSED METHOD

Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Feature selection is a term commonly used in data mining to describe the tools and techniques available for reducing inputs to a manageable size for processing and analysis. Fig. 2 shows a general framework of a clustering technique with and without feature selection on WEKA [10][11] and TANAGRA [12]. Fig. 3 shows the block diagram of steps of evaluation and comparison. In this experiment, clustering technique is applied on cardiology dataset using WEKA and TANAGRA tool which divides the dataset into clusters. Feature Selection select the important attributes from the dataset and then clustering technique is applied on the selected attributes and result is evaluated and compared. The performance of tools are compared based on accuracy for Cardiology dataset.

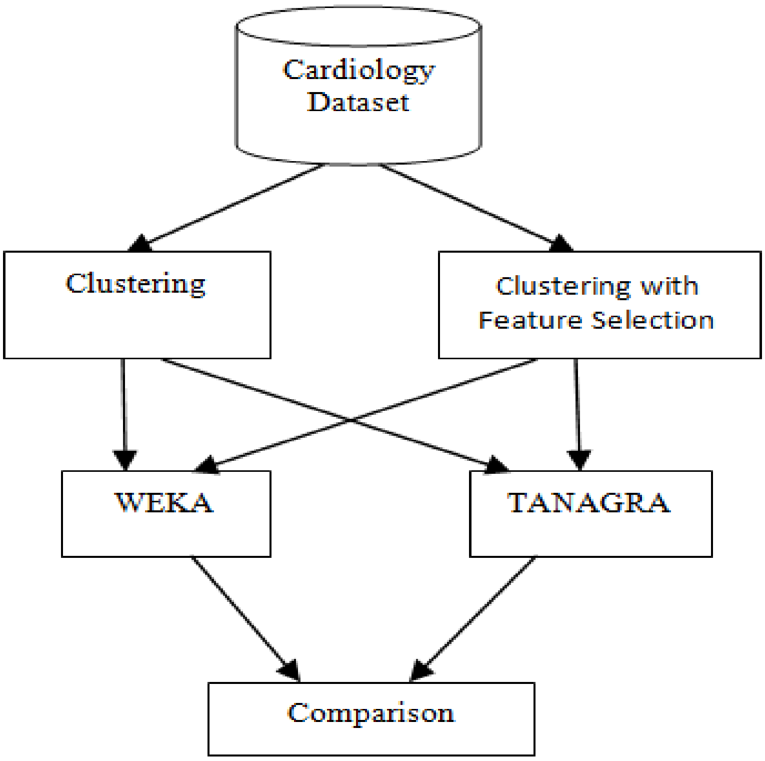


Fig 2 System Architecture

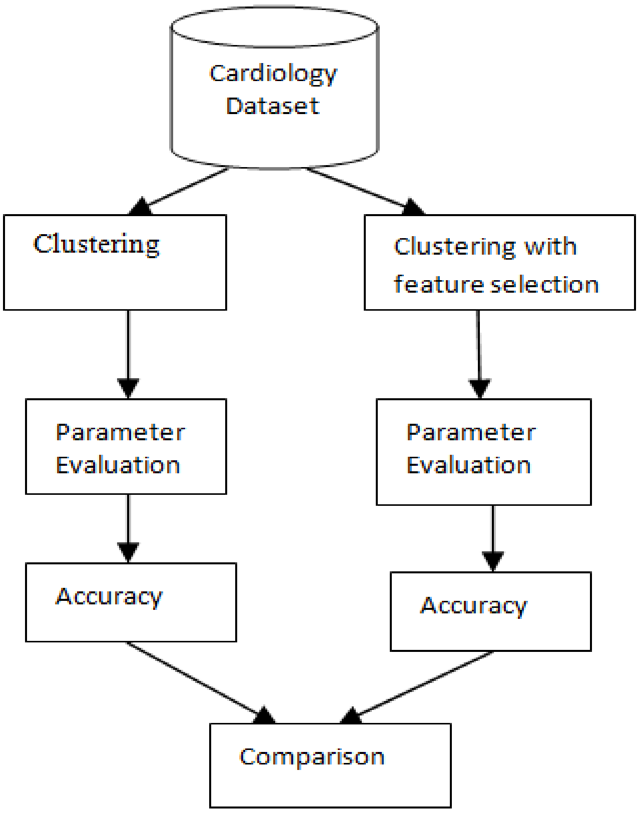


Fig 3 Block Diagram

A. Cardiology Dataset

Cardiology data set is downloaded from UCI machine learning website [13]. This dataset contains 303 instances and 14 attributes as nine are continuous attributes and five are discrete attributes. There are 138 samples in the dataset that are assigned to sick class and the other 165 samples are assigned to healthy class. Complete description of variables is shown in table 1.

TABLE 1
COMPLETE DESCRIPTION OF DATASET

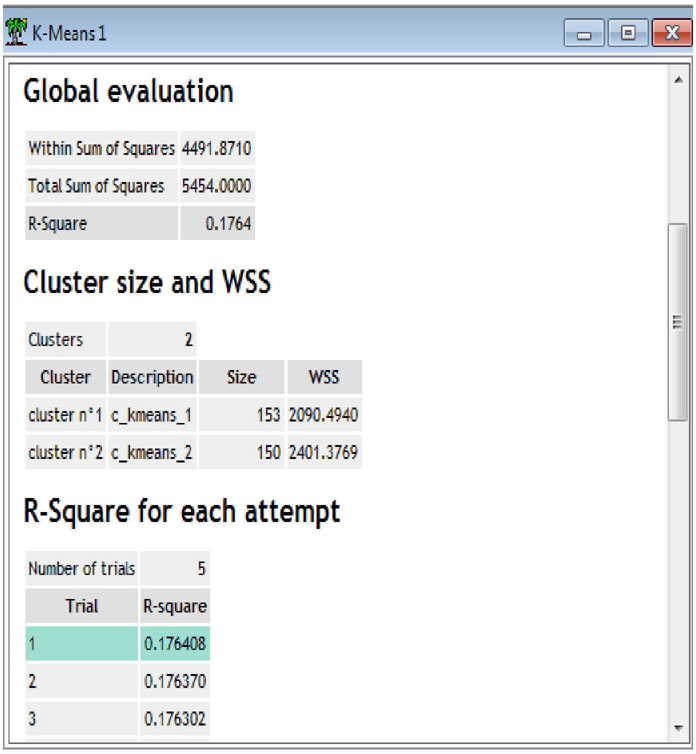
Dataset Description		
14 attributes		
303 examples		
Attribute	**Category**	**Information**
Age	Continue	-
Sex	Discrete	2 value
chest pain type	Discrete	4 value
blood pressure	Continue	-
Cholesterol	Continue	-
fasting blood sugar	Discrete	2 value
resting ecg	Discrete	3 value
maximum heart rate	Continue	-
Angina	Discrete	2 value
Peak	Continue	-
Slope	Discrete	3 value
colored vessels	Continue	-
Thal	Discrete	3 value
Class	Discrete	2 value

V. EXPERIMENTS RESULTS AND PERFORMANCE EVALUATION

In this experiment we present a comparative study K-Means algorithms with and without feature selection on WEKA and TANAGRA using accuracy and sum of squared error as a parameters used for comparison. The database used for this experiment is collected from UCI Repository of machine learning database. It consist of 303 instances and 14 attributes. The cardiology dataset for the experiment was first converted Excel Format i.e. .xls file.

A. Clustering with and without feature selection on TANAGRA

In order to perform experiment using TANAGRA, the file format for cardiology database has been changed to .txt file. Clustering in TANAGRA can be applied only on continuous values. But this dataset consisted of mixture of discrete and continuous values. In this experiment we have converted values of different attributes of the medical database in an appropriate binary format to perform desired experiment. A binary variable has only two states: 0 or 1, where 0 means that the variable is absent, and 1 means that it is present.



Global evaluation

Within Sum of Squares	4491.8710
Total Sum of Squares	5454.0000
R-Square	0.1764

Cluster size and WSS

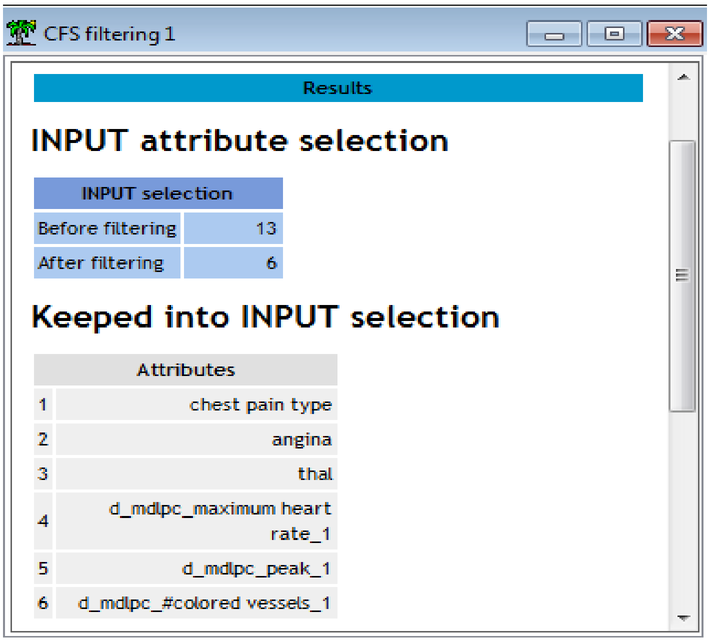
Clusters	2		
Cluster	Description	Size	WSS
cluster n°1	c_kmeans_1	153	2090.4940
cluster n°2	c_kmeans_2	150	2401.3769

R-Square for each attempt

Number of trials	5
Trial	R-square
1	0.176408
2	0.176370
3	0.176302

Fig. 4 clustering result of K-means method without feature selection

In this experiment K-means without feature selection gives accuracy 17.64%. Now, convert all the continuous attributes in to discrete using MDLPC from ‘Feature construction’ tab. Now we have all the attributes are of discrete type. Apply ‘CFS filtering’ from feature selection tab. This filtering technique selects 6 attributes out of 13 as shown in fig.5. Attributes like age, sex, blood pressure, and slope are not relevant for the diagnosis of heart disease. Therefore, these attributes have removed from the dataset. To perform K-means clustering with feature selection, transform the attributes in to binary continuous values using ‘0_1 Binarize’ from feature construction tab. Now apply K-means clustering on selected attributes. It is observed that accuracy of K-means method on cardiology dataset has increased to 27.69%.



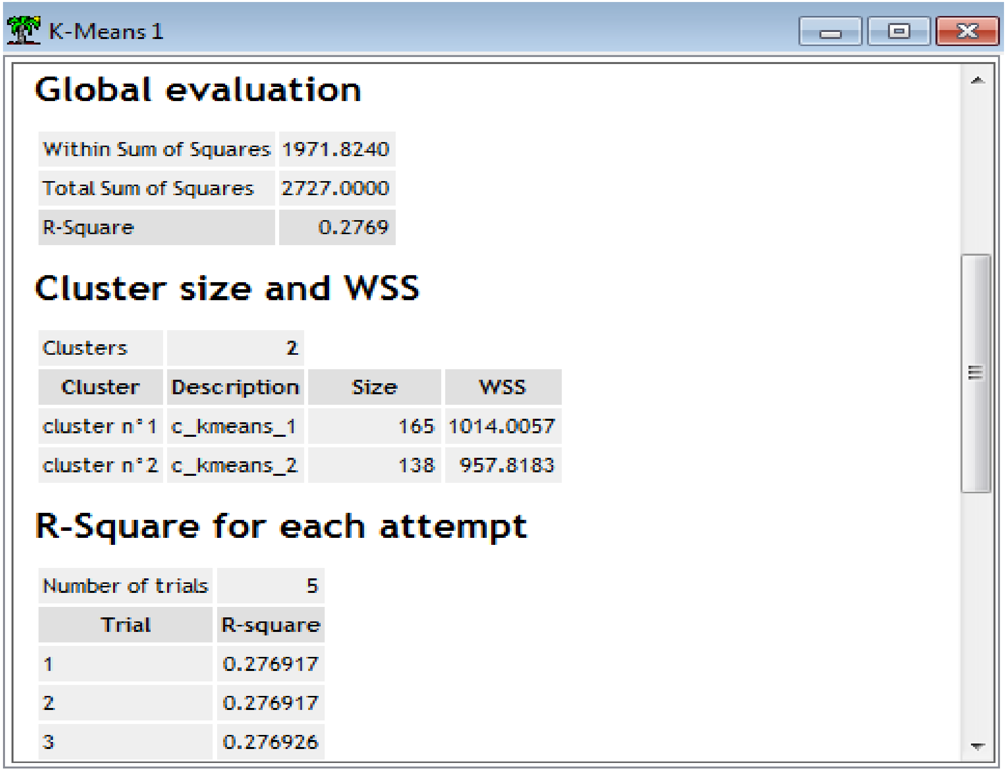
INPUT attribute selection

INPUT selection	
Before filtering	13
After filtering	6

Keeped into INPUT selection

	Attributes
1	chest pain type
2	angina
3	thal
4	d_mdlpc_maximum heart rate_1
5	d_mdlpc_peak_1
6	d_mdlpc_#colored vessels_1

Fig. 5 Features selected using filtering technique in TANAGRA



Global evaluation

Within Sum of Squares	1971.8240
Total Sum of Squares	2727.0000
R-Square	0.2769

Cluster size and WSS

Clusters	2		
Cluster	**Description**	**Size**	**WSS**
cluster n°1	c_kmeans_1	165	1014.0057
cluster n°2	c_kmeans_2	138	957.8183

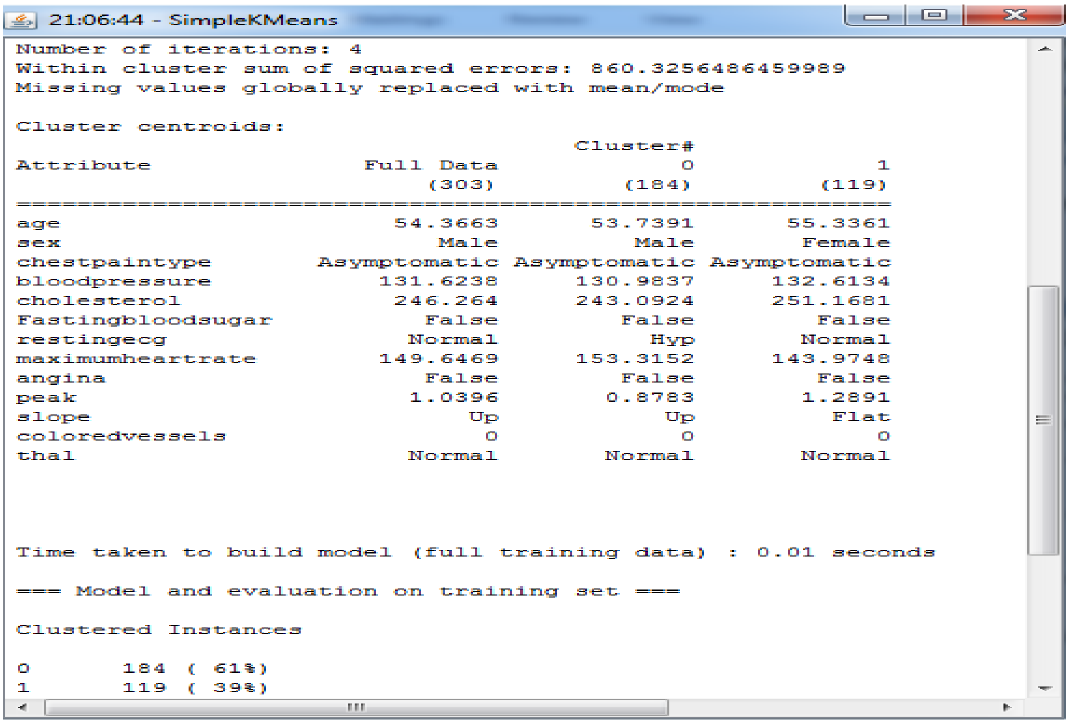
R-Square for each attempt

Number of trials	5
Trial	**R-square**
1	0.276917
2	0.276917
3	0.276926

Fig. 6 K-means clustering with feature selection

B. Clustering with and without feature selection on weka

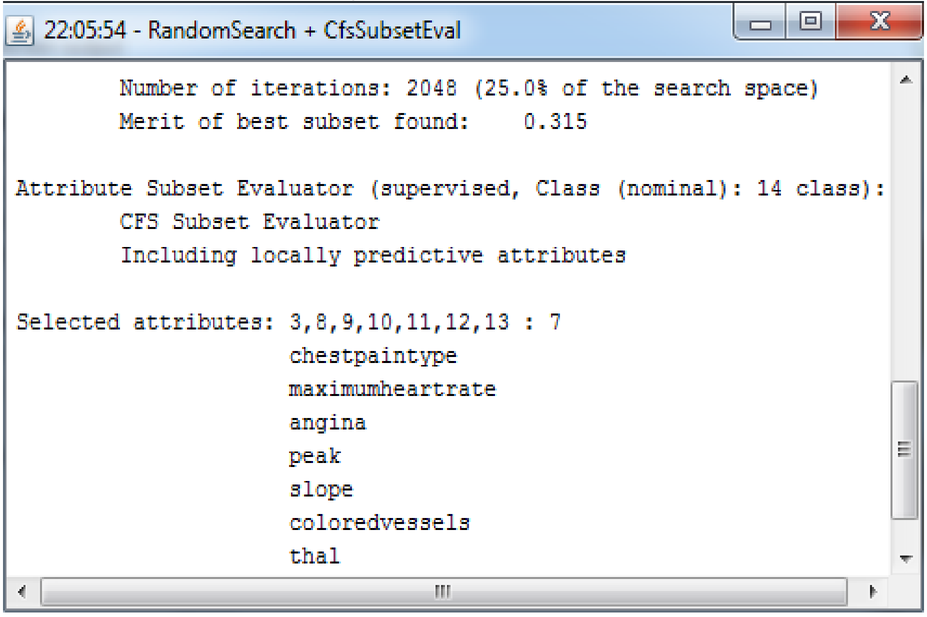
In this experiment we have performed K-means clustering on cardiology data set. This data set consists of both numeric and nominal values. Distance measure used for this experiment is Euclidean distance. Weka Simple K-means algorithm automatically handles a mixture of categorical and numerical attributes. We doing distance computations, where distances between categorical are assigned to 1 when are categorical values are different and to 0 when equal. Within cluster sum of squared error is 860.325. It divides 184 instances in cluster 1 and 119 in cluster 2. Values in cluster 1 are assigned to class "healthy" while values in cluster 2 are assigned to class "sick". Accuracy to perform this experiment is 50.9%.



```
Number of iterations: 4
Within cluster sum of squared errors: 860.3256486459989
Missing values globally replaced with mean/mode

Cluster centroids:
                                     Cluster#
Attribute              Full Data            0            1
                           (303)        (184)        (119)
==========================================================
age                      54.3663      53.7391      55.3361
sex                         Male         Male       Female
chestpaintype       Asymptomatic Asymptomatic Asymptomatic
bloodpressure           131.6238     130.9837     132.6134
cholesterol              246.264     243.0924     251.1681
Fastingbloodsugar          False        False        False
restingecg                Normal          Hyp       Normal
maximumheartrate        149.6469     153.3152     143.9748
angina                     False        False        False
peak                      1.0396       0.8783       1.2891
slope                         Up           Up         Flat
coloredvessels                 0            0            0
thal                      Normal       Normal       Normal



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       184 ( 61%)
1       119 ( 39%)
```

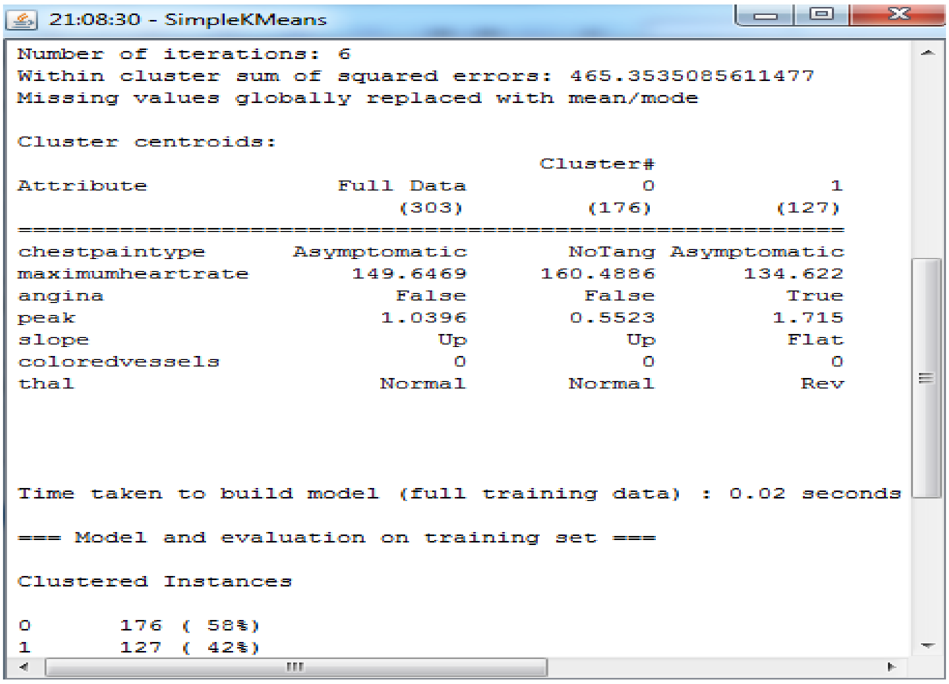
Fig. 7 K-means clustering method without feature selection

In this experiment we have performed cfsSubsetEval feature selection method over cardiology dataset to select relevant features. This method evaluates a subset of attributes which are more relevant for the diagnosis of heart disease. This method selects only 7 attributes (chest pain type, maximum heart rate, angina, peak, slope, colored vessels, thal) out for 13. After that, we have performed K-means clustering method on this subset. We have observed that with in cluster sum of square error has reduced to 465.35, and accuracy increased to 79.9%.



```
        Number of iterations: 2048 (25.0% of the search space)
        Merit of best subset found:    0.315

Attribute Subset Evaluator (supervised, Class (nominal): 14 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 3,8,9,10,11,12,13 : 7
                     chestpaintype
                     maximumheartrate
                     angina
                     peak
                     slope
                     coloredvessels
                     thal
```

Fig. 8 Features selected using filtering technique in TANAGRA



```
Number of iterations: 6
Within cluster sum of squared errors: 465.3535085611477
Missing values globally replaced with mean/mode

Cluster centroids:
                                    Cluster#
Attribute             Full Data            0            1
                          (303)        (176)        (127)
=========================================================
chestpaintype      Asymptomatic       NoTang Asymptomatic
maximumheartrate       149.6469     160.4886      134.622
angina                    False        False         True
peak                     1.0396       0.5523        1.715
slope                        Up           Up         Flat
coloredvessels                0            0            0
thal                     Normal       Normal          Rev



Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       176 ( 58%)
1       127 ( 42%)
```

Fig. 9 K-means method with feature selection on cardiology dataset

A. *Observations and Analyses*

It may be observed from Table 2 that accuracy of K-means clustering method is more in case of cardiology dataset on WEKA as compare to TANAGRA. WEKA Simple K-means algorithm automatically handles a mixture of categorical and numerical attributes. We have observed that accuracy of K-means clustering method is more with feature selection using WEKA tool on Cardiology dataset.

TABLE 2: ACCURACY OF K-MEANS CLUSTERING METHOD

Dataset used	**Parameters**		**Accuracy**	**SSE**
Cardiology	**Weka**	**K-means Without feature selection**	50.9%	860.325

		K-means With feature selection	79.9%	465.353
	Tanagra	**K-means Without feature selection**	17.64%	4491.8710
		K-means With feature selection	27.69%	1971.8240

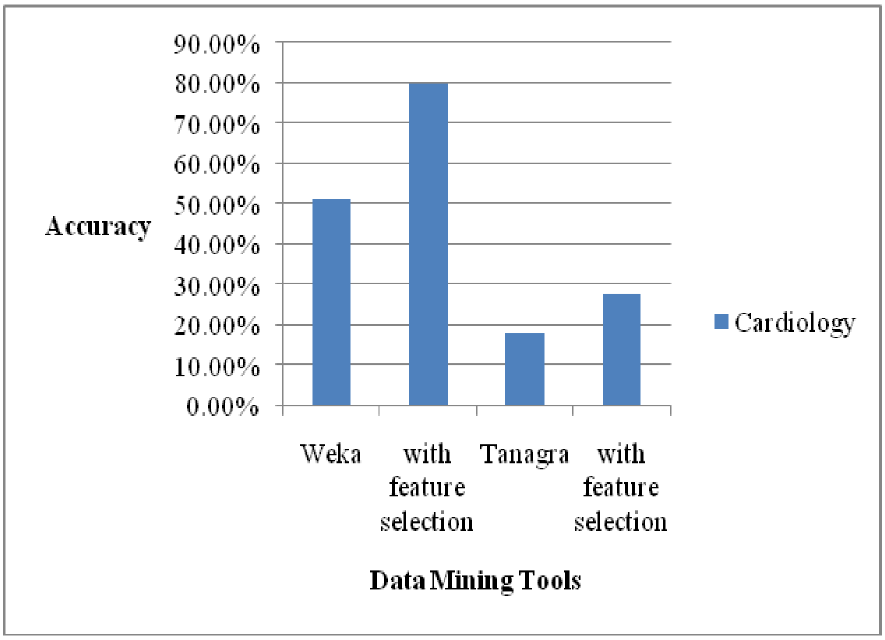


Fig. 10 Accuracy on Cardiology datasets using TANAGRA and WEKA

VI. CONCLUSION AND FUTURE SCOPE

This paper focuses on clustering algorithm such as K-means to discover useful knowledge from cardiology dataset. Clustering algorithms have been applied with and without feature selection on cardiology dataset using TANAGRA and WEKA tools. It has concluded that choice of a good feature can contribute a lot to clustering techniques. It can also be conclude that Weka gives better accuracy results with K-means as compared to TANAGRA. WEKA can handle a mixture of continuous and discrete values easily therefore accuracy of K-means clustering method is more, while clustering methods in Tanagra can be applied on continuous values. Therefore accuracy to perform clustering methods on discrete attributes is less using TANAGRA.

REFERENCES

[1] L. Kaufinan and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons,1990.

[2] YongSeog Kim, W. Nick Stree't, and Filippo Menczer, University of Iowa, USA, *Feature Selection in Data Mining*.

[3] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer Academic, 1998.

[4] Dash, M., & Liu, H. , "Feature Selection for Clustering". *Proc. of PAKDD-00*, pp. 110-121, 2000.

[5] P. Langley, "Selection of Relevant Features in Machine Learning," *Proc. AAAI Fall Symp. Relevance*, pp. 140-144, 1994.

[6] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection,"Artificial Intelligence, vol. 97, nos. 1-2, pp. 273-324, 1997.

[7] M. Dash, K. Choi, P. Scheuermann, and H. Liu, "Feature Selection for Clustering- a Filter Solution," Proc. Second Int'l Conf. Data Mining, pp.115-122, 2002.

[8] R.Caruana and D. Freitag, "Greedy Attribute Selection", Proc.11th Int'l Conf. Machine Learning, pp. 28-36, 1994.

[9] S. Das, "Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection," Proc. 18th Int'l Conf. Machine Learning, pp. 74-81, 2001.

[10] Holmes, G., Donkin,A., Witten,I.H., "WEKA a machine learning workbench,". In: Proceeding second Australia and New Zeeland Conference on Intelligent Information System, Brisbane , Australia, pp.357- 361, 1994.

[11] Weka 3- Data Mining with open source machine learning software available from :- http://www.cs.waikato. ac.nz/ml/ weka/.

[12] Tanagra – free data mining software for teaching and research [Online]. Available: http://eric.univ- lyon2.fr/_ricco/tanagra/en/tanagra.

[13] UCI Machine Learning Repository. Irvine, CA: University of California, Center for Machine Learning and Intelligent Systems. [Online]. Available:http://archive.ics.uci.edu/ml/.