Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
CASE STUDY
Available Online at www.ijarcs.info

Clustering Algorithms on Diabetes Data: Comparative Case Study

Usha G Biradar
Assistant Professor
P. G. Department of Computer Science, Alva's College
Moodbidri -India

Dr. Deepa S Mugali
Post Graduate Student
Department of Opththlmology Mahadevappa Rampure
Medical College(MRMC), Kalaburgi-India

Abstract: Data Clustering is used to extract meaningful information and plays a vital role in data mining. Its main job is to group the similar data together based on the characteristic they possess. This paper represents the performance of three clustering algorithms such as Hierarchical clustering, EM and K Means clustering algorithm.
The Diabetes dataset is used for the comparison of those clustering algorithms based on the performance. This comparative study focuses on use of data mining tool to analyze a previously obtained data set using Weka and Tanagra. The results were compared to find algorithm yields and best result presented.

Keywords: Data Mining, Cluster techniques, Hierarchical clustering, EM and K Means clustering algorithm, Weka, Tanagra.

I . INTRODUCTION

Data mining is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, data warehouses, or other information repositories. It is a young interdisciplinary field, drawing from areas such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing. Other contributing areas include neural networks, pattern recognition, spatial data analysis, image databases, signal processing, and many application fields, such as business, economics, and bioinformatics.

Data clustering is a process of extracting previously unknown, valid, positional useful and hidden patterns from large data sets.

In supervised learning, the hierarchical clustering is one of the most frequently used methods and it is typically more effective in detecting the true clustering structure of a data set than partitioning algorithms.[1]

In computer vision community, the k-eans algorithm is one of the most commonly used clustering algorithms which can be used for its simplicity and effectiveness. It's an iterative algorithm in which, each iteration new cluster centers are computed and each data point is re-assigned to its nearest center. And also the k-means clustering algorithm is widely used in machine learning for clustering and quantization. The EM algorithm alternates between maximizing F with Respect to Q(theta fixed) and then maximizing F with respect to theta(Q fixed).[2]

II. DATA MINING TASKS

Generally different classes of tasks can be achieved by exercising DM .

a. Prediction: this task aims at forecasting what might happen in the future by estimating the likelihood of a certain event's occurrence.

b. Classification: it is usually exercised to identify group membership in a population instances. Popular classification techniques use Neural Networks (NN) and Decision Trees.

c. Clustering: it is applied to position elements of a database into specific groups according to some attributes. The most frequently modi operandi are k-means and expectation maximization.

d. Association: this area of DM aims at analyzing data to identify consolidated occurrence of events and uses the criteria of support and confidence. It is known to be applied in customer behavior and machine learning. A popular procedure used is the Apriori algorithm.

e. Sequential Analysis: this task targets the occurrence of special sequence of events where time plays a key role. It leads to the Identification of the events that most likely will lead to later ones with a specified minimum support or percentage.

III. DATASET DESCRIPTION

The diabetes dataset is collected from the UCI repository and it contains 769 instances and 9 attributes. Diabetes mellitus is a group of metabolic diseases characterized by high blood sugar (glucose) levels that result from defects in insulin secretion. Patients with high blood sugar will typically experience polyuria, they will become increasingly thirsty and hungry [3]. Glucose is vital to human health, because it's an important source of energy for the cells that make up the muscles and tissues and also the brain's main source of fuel [4].

IV. TOOLS DESCRIPTION

A. WEKA

WEKA toolkit [5] is a widely used toolkit for machine learning and data mining that was originally developed at the University of Waikato in New Zealand. It contains a large collection of state-of-the-art machine learning and data mining algorithms written in Java. WEKA contains tools for

regression, classification, clustering, association rules, visualization, and data pre-processing. WEKA has become very popular with the academic and industrial researchers, and is also widely used for teaching purposes.

B. TANAGRA

Tanagra is free data mining software for academic and research purposes. It offers several data mining methods like exploratory data analysis, statistical learning and machine learning. The first purpose of the Tanagra project is to give researchers and students easy-to-use data mining software. The second purpose of TANAGRA is to propose to researchers an architecture allowing them to easily add their own data mining methods, to compare their performances. The third and last purpose is that novice developers should take advantage of the free access to source code, to look how this sort of software was built, the problems to avoid, the main steps of the project, and which tools and code libraries to use for. In this way, Tanagra can be considered as a pedagogical tool for learning programming techniques as well [6].

V. COMPARATIVE STUDY

We compare result Weka and Tanagra tool. In this study measures are calculated by using the performance factors such as the clustering accuracy and execution time. And also the comparative analysis for the Diabetes datasets is performed to predict
the finest algorithm. The accuracy measure and the execution time for the clustering algorithms are depicted in Table 1.

TABLE 1

Data Mining Tool	Clustering Algorithm	No. of clusters	Cluster (0)	Cluster (1)	Cluster 0 (%)	Cluster 1 (%)	Time (second)
WEKA	EM	2	432	336	56	44	0.32
	KMEANS	2	500	268	65	35	0.02
	HIRARCHICAL	2	268	500	35	65	2.34
TANAGRA	EM	2	408	360	53	47	0.06
	KMEANS	2	493	275	65	35	0.03
	HIRARCHICAL	2	342	426	45	55	3.36

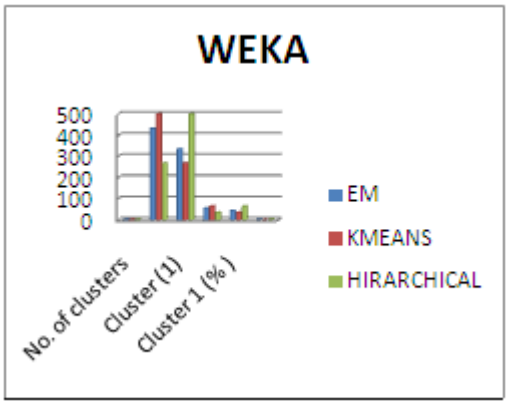

Figure 1

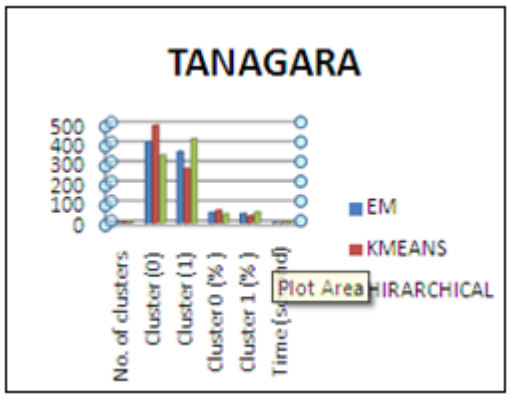

Figure 2

The performance measures for the clustering algorithms is graphically represented in the Figure 1 and 2.

The study was carried out to the diabetes datasets by using the training set . From the results it is inferred that k means clustering algorithm performs better as compared to other clustering algorithms in both tool. The k means algorithm gives
more correctly clustered instances compared to others in both tool.

VI CONCLUSION AND FUTURE WORK

In this paper the performance is evaluated for the clustering .The algorithms are analyzed by using the trained set. The performance measures are analyzed based on the number of clustered instances and the execution time taken for clustering the instances. From the study results it is inferred that the K means algorithm gives better performance when comparing with the other two algorithms by using the Diabetes dataset in both tool.

In future the clustering algorithms can be experimented on other datasets also. And in future the k means clustering algorithm will modify to obtain more effective results. And also the k means clustering algorithm can be analyzed using various parameters such as the cross validation, percentage split, and supplied test set.

VII. REFERENCES

[1] R.Nithya1, P.Manikandan2, Dr.D.Ramyachitra, Analysis of clustering technique for the diabetes dataset using the training set parameter, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 9, September 2015

[2] Murlidher Mourya An Effective Execution of Diabetes Dataset Using WEKA (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (5) , 2013, 681-682

[3] http://www.medicinenet.com/diabetes_mellitus/page2.htm#what_is _diabetes.

[4] http://www.mayoclinic.org/diseasesconditions/diabetes/basics/definition/con-20033091

[5] WEKA, the University of Waikato, Available at: http://www.cs.waikato.ac.nz/ml/weka/, (Accessed 20 April 2011).

[6] Tanagra – a Free Data Mining Software for Teaching and Research, Available at: http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html, (Accessed 20 April 2011).