ISSN:0975-9646

Murlidher Mourya et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (5) , 2013, 681-682

An Effective Execution of Diabetes Dataset Using WEKA

Murlidher Mourya , Phani Prasad

Department of CSE, Vardhaman Engineering College,
Hyderabad,India.

***Abstract*— Mining of the Data now days plays a major role and concern in the present world in the industry and also in the research areas. Many tools and application software's are evolved to mine the data and to present the best data visualization. One of the Data Mining tool available is WEKA, in which contain many machine level and other types of algorithms. In the present paper the data classification is a medical dataset of diabetes category in which we cluster the dataset using various clustering algorithms like EM, k-means, OPTICS and the results are depicted.**

***Keywords*— Dataset, Data Mining, Clustering, Weka.**

INTRODUCTION

Data Mining is the extraction and retrieval of useful and unknown data in the past. Data Mining also involves the retrieval and analysis of data that is stored in a Data ware house. Some of the major techniques of Data Mining are Classification, Association, Clustering and regression etc. in this paper our main goal is to show the cluster analysis on the diabetes dataset .

Clustering is nothing but grouping of set of objects of similar type in to one group or category and set of dissimilar objects as other category. Clustering is also referred as the explorative data mining task which is used in areas like image processing and image analysis, information retrieval, pattern recognition and Bioinformatics. The purpose of using WEKA tool is because of its user friendliness and it is the best mining tool for analyzing data sets along with languages like java.

I. DATASET

A dataset (or data set) is a collection of data. Most commonly a dataset corresponds to the contents of a single database table, or a single statistical data matrix, where each column of the table represents a particular variable, and each row corresponds to a given member of the dataset in question. The dataset lists values for each of the variables, such as height and weight of an object, for each member of the dataset. Each value is known as a datum. The dataset may comprise data for one or more members, corresponding to the number of rows

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

II. DATA MINING AND WEKA ENGINE

WEKA is a collection of various machine algorithms for Data Mining tasks. The algorithms can be directly applied to the data set or they can be taken through java code. WEKA contains tools for Pre-processing, Classification, Clustering, Association, Data visualization. WEKA is developed by University of Waikato.

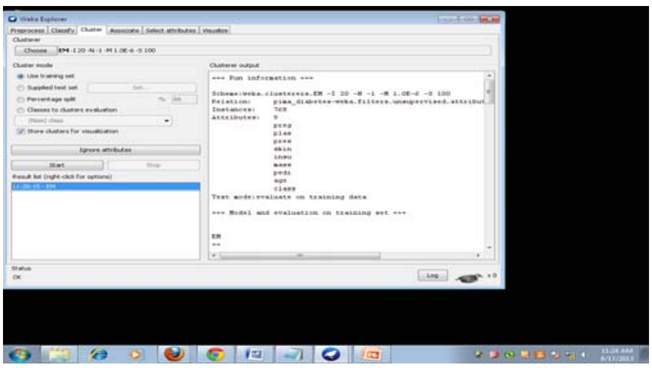
III. CLUSTERING

Cluster analysis itself is not one specific algorithm but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

In the present paper we are dealing with Diabetes Data set with the attributes as "Preg", " plas", "pres", "skin", "age", "insu", "pres", "mass", "class". With the help of these attributes we cluster the data set using WEKA tool.

IV. EM ALGORITHM AND CLUSTERING

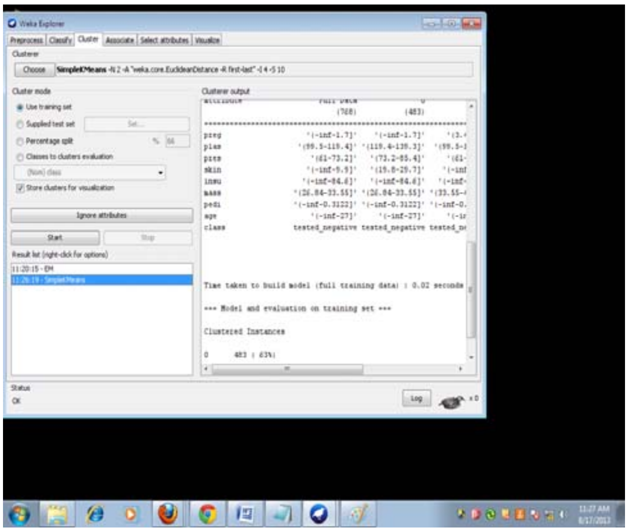
The EM algorithm alternates between maximizing F with Respect to Q(theta fixed) and then maximizing F with respect to theta(Q fixed)



V. CLUSTERING BY K-MEANS ALGORITHM

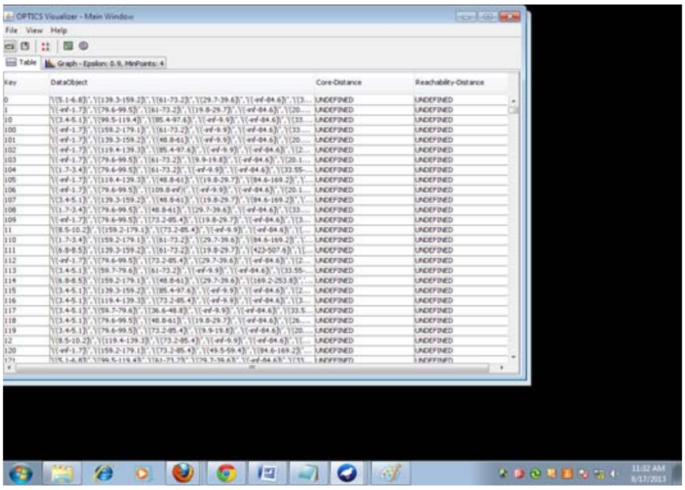
K-MEANS CLUSTERING

Is a method of vector quantization originally from signal processing , that is popular for cluster analysis in data mining. K- means clustering aims to partition N observation into K clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster , this results in a partition of the data space into voroni cells.



VI. CLUSTERING BY OPTICS

Ordering points to identify the clustering structure (**OPTICS**) is an algorithm for finding density-based clusters in spatial data. Its basic idea is similar to DBSCAN, but it addresses one of DBSCAN's major weaknesses: the problem of detecting meaningful clusters in data of varying density. In order to do so, the points of the database are (linearly) ordered such that points which are spatially closest become neighbors in the ordering. Additionally, a special distance is stored for each point that represents the density that needs to be accepted for a cluster in order to have both points belong to the same cluster. This is represented as a dendrogram



VII. CONCLUSIONS

With the Help of this WEKA tool effective and efficient execution of the Diabetes data set has been done and in future we can extend this work by using other techniques like classification, Association rules etc . not only for this dataset but to any other data sets also.

REFERENCES

[1] Mats Jontell, Oral medicine, SahlgrenskaAcademy, Göteborg University (1998) “AComputerised Teaching Aid in Oral Medicine andOral Pathology. “ Olof Torgersson, department ofComputing Science, Chalmers University ofTechnology, Göteborg.

[2] T. Mitchell, "Decision Tree Learning", in T.Mitchell, Machine Learning (1997) the McGraw-Hill Companies, Inc., pp. 52-78.

[3] P.Yashoda, M.Kanan, Analysis of a population of diabetic patients databases in WEKA tool, IJSER, vol2, issue5, may 2011.

[4]. Witten Ian H., E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Ch. 8, © 2000 Morgan Kaufmann publishers.

[5]. http://www.cs.waikato.ac.nz/ml/weka/, accessed 06/05/21.