DIABETES PREDICTION USING MULTIPLE CLASSIFICATION ALGORITHMS

THOTA YAMINI KALYANI [#1], K.VENKATESH [#2], D.D.D.SURIBABU [#3]

[#1] MCA Student, Master of Computer Applications,

D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.

[#2] Assistant Professor, Master of Computer Applications,

D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.

[#2] Head & Associate Professor, Department of CSE,

D.N.R. College of Engineering and Technology, Bhimavaram, AP, India.

ABSTRACT

The effectiveness of a particular algorithm may be influenced by a number of different factors. However, the influence of a particular factor like quality can be considered to identify the effectiveness of an algorithm for a given data source. The project provides a comparative analysis of two clustering algorithms namely K-Means, and MST applied to diabetes dataset. Besides rapidly generating the clusters, the analysis also provides a basis for determining the quality of the clusters generated and helps in identifying the algorithm that generates good quality clusters. As Data Characterization is a summarization of the general characteristics or features of a target class of data, the characteristics of diabetes data are also analyzed taking into account positively tested records as target class of data using the approach of Attribute Oriented Induction.

Key Words:

Characterization, Summarization, Records, K-Means, Minimum Spanning Tree.

I. INTRODUCTION

The fundamental explanation that information mining has pulled in a lot of consideration in the data business lately is because of the wide accessibility of immense measures of information and the impending requirement for transforming such information into helpful data and information. The data and information picked up can be utilized for applications extending from business the executives, creation control and market examination to building plan and science investigation.

The consistent and stunning advancement of PC equipment innovation from the previous three decades has let to enormous supplies of ground-breaking and reasonable PCs, information assortment gear and

capacity media. This innovation gives an extraordinary lift to the database and data industry and makes countless databases and data storehouses accessible for exchange the board, data recovery and information investigation.

Among the zones of information mining, Cluster Analysis has gotten a lot of consideration. Grouping investigates information objects without speaking with a realized class name. As a rule, the class marks are absent in the preparation information just in light of the fact that they are not known in the first place. Grouping can be utilized to produce such marks. The articles are bunched or gathered dependent on the rule of amplifying the intra group comparability and limiting the bury bunch closeness. That is, the groups of items are shaped so protests inside a bunch have high comparability in contrast with each other, yet are extremely unlike articles in another bunch. Each group that is framed can be seen as a class of articles, from which rules can be determined. Bunching can likewise encourage scientific categorization arrangement, that is, the association of perceptions into a pecking order of classes that bunch comparable occasions together. There are various uses of information mining, which fit into this system.

The general execution of grouping is subject to the disclosure of value bunches. In this manner a greater part of calculations are worried about proficiently deciding the arrangement of groups in a given exchange or social database. The issue is basically to produce the arrangement of groups in the database. Consequently, various calculations were presented which target producing quality groups. These calculations contrast from each other in the strategy for creating the groups.

This task ―Quality Analysis and Characteristic Evaluation of Diabetes Data utilizing Clustering Techniques‖ is worried about gathering up of information to such an extent that the intra group likeness is amplified and the entomb bunch similitude is limited. That is, information questions that are comparable, fall in a similar bunch and are not at all like articles in different groups. As execution of bunching is reliant on the age of value groups, the nature of the groups produced is resolved utilizing the Sum of Squared Error approach and the best bunching calculation that creates bunches of good quality is considered for portrayal of diabetes information.

II. LITERATURE SURVEY

In this section we will mainly discuss about the background work that is carried out in order to prove the performance of our proposed Method. Now let us discuss about them in detail

MOTIVATION

Grouping is a difficult field of exploration where its potential applications represent their own unique prerequisites. Coming up next are commonplace prerequisites of bunching in information mining:

• Scalability: Many bunching calculations function admirably on little informational collections containing less than 200 information objects; notwithstanding, an enormous database may contain a great many articles. Grouping on an example of a given enormous informational collection may prompt one-sided results. Profoundly adaptable grouping calculations are required.

• Ability to manage various kinds of properties: Many calculations are intended to group span based (numerical) information. Be that as it may, applications may require grouping different kinds of information, for example, twofold, downright (ostensible), and ordinal information, or blends of these information types.

• Discovery of bunches with self-assertive shape: Many grouping calculations decide groups dependent on Euclidean or Manhattan separation measures. Calculations dependent on such separation estimates will in general find round bunches with comparable size and thickness. In any case, a bunch could be of any shape. It is imperative to create calculations that can identify groups of self-assertive shape.

• Minimal prerequisites for area information to decide input boundaries: Many grouping calculations expect clients to enter certain boundaries in bunch, (for example, the quantity of wanted groups). The grouping results can be very delicate to enter boundaries.

Boundaries are frequently difficult to decide, particularly for informational indexes containing high dimensional items. This weights clients, yet in addition makes the nature of bunching hard to control.

• Ability to manage loud information: Most genuine databases contain anomalies or missing, obscure, or wrong information. Some bunching calculations are touchy to such information and may prompt groups of low quality.

• Insensitive to the request for input records: Some bunching calculations are touchy to the request for input information: for instance, a similar arrangement of information, when given various orderings to such a calculation, may produce significantly various groups. It is imperative to create calculations that are harsh toward the request for input.

• High dimensionality: A database or an information distribution center can contain a few measurements or traits. Many grouping calculations are acceptable at dealing with low-dimensional information, including just a few measurements. Natural eyes are acceptable at making a decision about the nature of grouping for up to three measurements. It is trying to bunch information objects in high-dimensional space, particularly thinking about that such information can be meager and profoundly slanted.

• Constraint-based bunching: Real-world applications may need to perform grouping under different sorts of requirements. Assume that our main responsibility is to pick the areas for a given number of new programmed money apportioning machines in a city. To settle on this, we may bunch house holds while considering the limitations, for example, city's waterways and thruway systems, and client prerequisites per district. A difficult assignment is to discover gatherings of information with great grouping conduct that fulfill indicated imperatives.

• Interpretability and Usability: Users anticipate that grouping results should be interpretable, intelligible and usable. That is, grouping may should be tied up with explicit semantic understandings and applications. It is critical to concentrate how an application objective may impact the choice of grouping strategies.

III. EXISTING METHODOLOGY

In the existing system there is no single application which can compare a set of data mining algorithms in order to identify the performance of diabetes patients and hence we cannot able to get the accurate analysis about these algorithms.

LIMITATIONS OF THE EXISTING METHODOLOGY

The following are the limitation of existing system. They are as follows:

1) There is no application which can integrate multiple algorithms to read the performance of diabetes dataset.
2) There is no application which can compare the time delay for identifying the performance of diabetes dataset.
3) No proper method which can show the comparative analysis of multiple data mining algorithms.

IV. PROPOSED METHODOLOGY

In this proposed application we try to construct a application by taking two data mining algorithms and check the importance of every individual algorithm in finding the diabetes dataset.

ADVANTAGES OF THE PROPOSED SYSTEM

The following are the advantages of the proposed system. They are as follows:

1) Here we used K-means algorithm and try to cluster the records which are having similar type of values.
2) We can also find the time delay for calculating the appropriate clusters using these algorithms.
3) By using MST algorithm we can able to identify the nearest neighbor records which are matched with the clustered sets and we can able to see the common factors for getting diabetes.
4) Here we can see the complexity of detecting the diabetes patients having similar qualities.

IV. IMPLEMENTATION STAGE

Implementation Stage is where the hypothetical structure is changed over into automatically way. In this stage we will partition the application into various modules and afterward coded for arrangement. The application is separated essentially into following 4 modules. They are as per the following:

1) Load Dataset Module

2) Normalization Module

3) Choose Algorithm Module

4) Report Generation with Quality Estimation module

Now let us discuss about each and every module and sub modules which are present in this application.

1) LOAD DATASET MODULE

Here we try to collect female patients diabetes dataset which contains 8 attributes to test the presence of diabetes. All the records are having some values assigned for the 8 attributes. For instance Diabetes database is as follows.

6 148 72 35 0 33.6 0.627 50
1 85 66 29 0 26.6 0.351 31
8 183 64 0 0 23.3 0.672 32
1 89 66 23 94 28.1 0.167 21
0 137 40 35 168 43.1 2.288 33

Where each row represents an item and the columns correspond to the attribute values. The above database is normalized with the Normalization algorithm and then provided as input to the clusters.

2) NORMALIZATION MODULE

This method of normalization is used when the actual minimum and maximum of attribute *A* are unknown, or when there are outliers that dominate the min-max normalization. Here we try to check all the attributes are containing values within the limit specified by the diabetes dataset. If there are any in-complete values this will be removed.

3) CHOOSE TYPE OF ALGORITHM MODULE

Here we try to choose the type of algorithm to check the quality of diabetes and here we try to use two algorithms like : K-Means and MST algorithms and try to check the overall quality of both algorithms.

4) REPORT GENERATION MODULE

Here the report is generated after executing both the algorithms and we try to check which algorithm is best by observing the quality of both the algorithms. From the comparison results we clearly state that MST is best compared with K means in finding the clustering.

V. EXPERIMENTAL REPORTS

K-MEANS ALGORITH REPORT

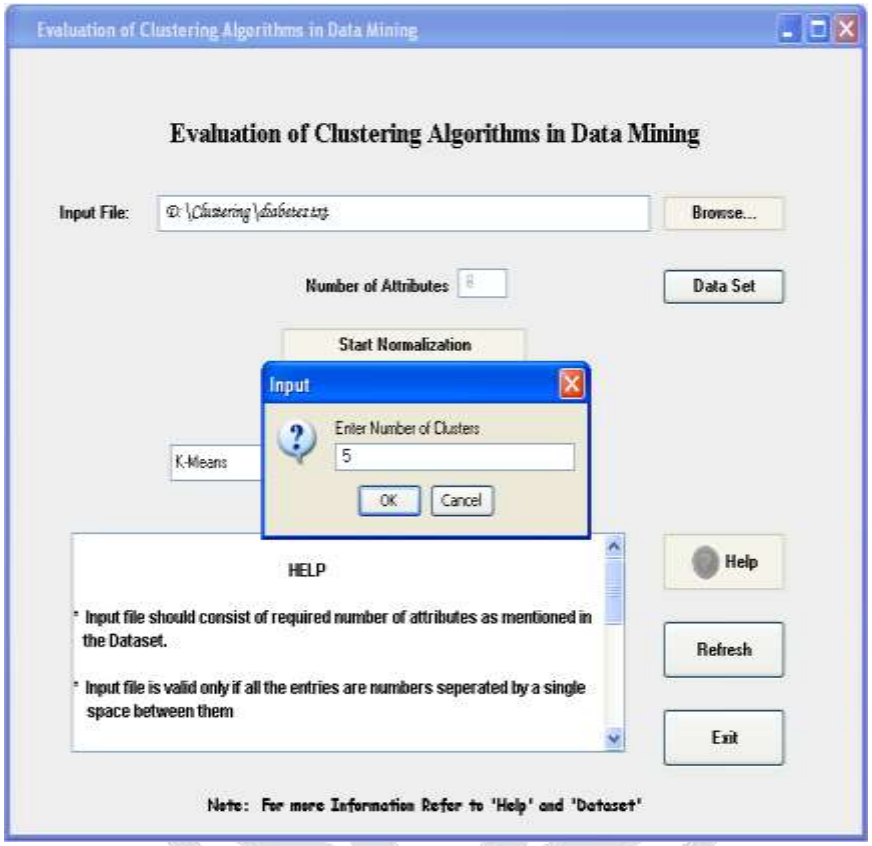

Figure . Represents the Number of Clusters Selection

K-MEANS OUTPUT

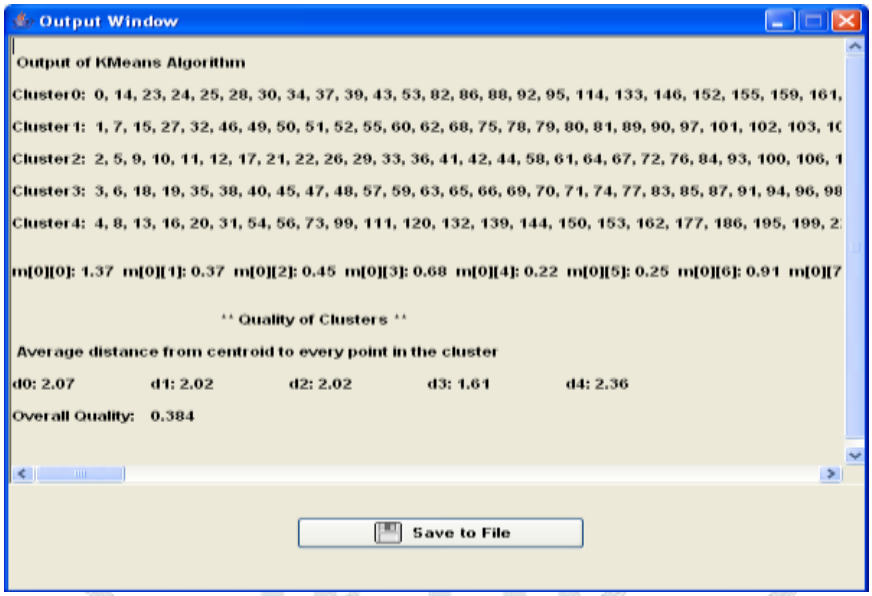

Figure. Represents the K-Means Output

REPRESENT THE MST REPORT

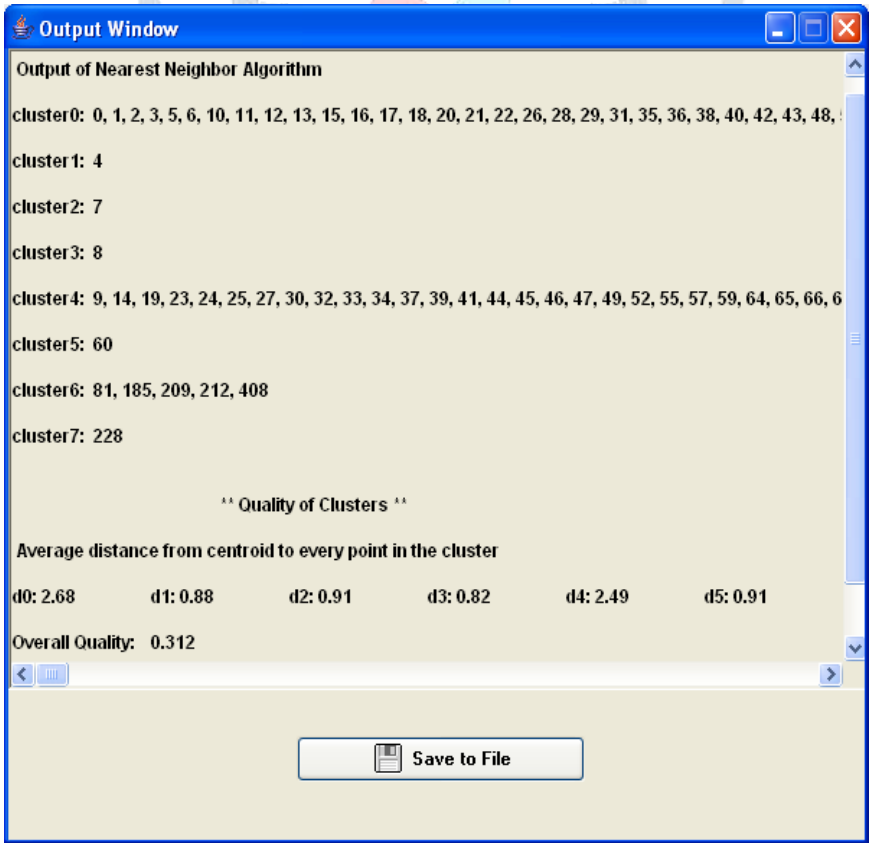

Figure .Represents the MST OUTPUT

VI. CONCLUSION

Clustering data is a canonical task, fundamental for many data mining and other applications. Especially by clustering, one can identify dense and sparse regions and, therefore, discover overall distribution patterns and interesting correlations among data attributes. Very much helpful in diagnosis of patients. Here we finally concluded that MST is giving good quality of clusters than compared with the primitive clustering algorithm like K-Means.

VII. REFERENCES

1. Jiawei Han, Micheline Kamber, Data Mining and Concepts.

2. Margaret H. Dunham, Data Mining and Introductory to Advanced Topics.

3. Timothy C Lethbridge & Langaniere, McGrawHill Co., Object Oriented Software Development using UML & Java.

4. Grady Booch, Ivan Jacobson, James Rumbaugh, Addison Wesley 1999, The Unified Modeling Language User Guide.

5. Roger. S. Pressman Ph.D., McGrawHill, Software Engineering A practitioner's Approach.

6. Herbert Schildt, The Complete Reference Java 2.

7. Diabetes data set. http://www.ncc.up.pt/liacc/ML/statlog/datasets/diabetes.dat

8. http://home.earthlink.net/~salhir/thefoundationoftheuml.html

9. Data mining: http://www.statsoft.com/textbook/stdatmin.html

10. Cluster Analysis: http://www.statsoft.com/textbook/stcluan.html

11. K-Means Algorithm:
http://cne.gmu.edu/modules/dau/stat/cluatalgs/clust5_frm.html

12. Minimum spanning tree method:
http://cne.gmu.edu/modules/dau/stat/clustgalgs/clust4_frm.html