Article can be accessed online at http://www.publishingindia.com

Clustering and Classifying Diabetic Data Sets Using K-Means Algorithm

M. Kothainayaki*, P. Thangaraj**

Abstract

The k-means algorithm is well known for its efficiency in clustering large data sets. However, working only on numeric values prohibits it from being used to cluster real world data containing categorical values. In this paper we present the Classification of diabetic's data set and the k-means algorithm to categorical domains. Before classify the data set preprocessing of data set is done to remove the noise in the data set. We use the missing value algorithm to replace the null values in the data set. This algorithm is also used to improve the classification rate and cluster the data set using two attributes namely plasma and pregnancy attribute.

Keywords: Classification, Cluster Analysis, Clustering Algorithms, Categorical Data, Pre-processing

1. Introduction

Classification is a mechanism to classify the data set and name the classes. After classification calculate the classification rate using the formula. Using this algorithm the data set is classified into two class label namely tested_positive and tested_negative. The data set is containing nine attributes namely preg, plas, mass, age, insu, skin, pedi, pres and class.

Partitioning a set of objects in databases into *homogeneous groups* or *clusters* is a fundamental operation in data mining. Clustering is a popular approach to implementing the partitioning operation. Clustering methods partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria.

The data sets to be mined often contain millions of objects described by tens, hundreds or even thousands of various types of attributes or variables (interval, ratio, binary, ordinal, nominal, etc.). This requires the data mining operations and algorithms to be scalable and capable of dealing with different types of attributes.

In this paper we present algorithms that use to classify the data set into two classes and compare with standard. The *k*-means to cluster data having categorical values.

2. Literature Review

A lot of research work has been done on various medical data sets including Pima Indian diabetes dataset. The authors [6] has implemented their algorithm and achieved the accuracy in classifying and clustering the diabetics datasets. In their experiment, they eliminated Incorrect labeled instance by using K-means clustering followed by feature extraction using GA_CFS. The resultant dataset is divided into training data and test data using 60-40 ratio. Experiments were carried out for different values of k ranging k from 1 to 15. The accuracy Diabetic data set using proposed method without feature selection is 95.56% with k = 5.

D. Vijayalakshmi & K. Thilagavathi has analysis that the clustering algorithm based on a graph b-coloring technique was used to cluster Pima Indian diabetic dataset. They implemented, performed experiments, and compared with KNN Classification and K-means clustering. The results show that the clustering based on

* Assistant Professor, Department of Computer Applications, Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India. E-mail: kothaimk@gmail.com

** Professor & Head, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India.

graph coloring out performance than the other clustering approach in terms of accuracy and purity.

The main purpose of the Diabetic Patients Databases system [3] is to guide diabetic patients during the disease. Diabetic patients could benefit from the diabetes expert system by entering their daily glucoses rate and insulin dosages; producing a graph from insulin history; consulting their insulin dosage for next day. The diabetes expert system is not only for diabetic patient, but also for the people who suspect if they are diabetic. It's also tried to determine an estimation method to predict glucose rate in blood which indicates diabetes risk.

3. Notation

We assume that in a database objects from the same domain are represented by the same set of attributes, *A*1, *A*2,….*Am*. Each attribute *Ai* describes a domain of values, denoted by *DOM(Ai)*, associated with a defined semantic and data type. Different definitions of data types are used in data representation in databases and in data analysis.

A numeric domain is represented by continuous values. A domain *DOM(Aj)* is defined as categorical if it is finite and unordered. A special value, denoted by €, is defined on all categorical domains and used to represent missing values.

It means the two objects have equal values for the attributes *A*1,*A*2,…,*Am*. For example, two patients in a data set may have equal values for the attributes Age, Sex, Disease and Treatment.

4. The Classification Algorithm

The classification algorithm is used to classify the data set and named the class label. Before classification, the data are preprocessed to remove the null values. We used the missing values algorithm to remove the null values. Instance of null values, replace into the mean value of each attribute. The algorithm is checked the plasma level and segregate the class with sequence of condition like age is less than 27 and mass is less than 37 etc. the attributes are declared and retrieved from the database.

Using this algorithm, all the instances are classified into two class label namely tested_positive and tested_negative. In this algorithm tested using the 20 sample data and classification is achieved for that sample data.

5. The *k*-means Algorithm

The *k*-means algorithm is the mostly used clustering algorithms, is classified as a *partitional* or *nonhierarchical* clustering method. Given a set of numeric objects *X* and an integer number *k* (≤*n*),the *k*-means algorithm searches for a partition of *X* into *k* clusters that minimizes the within groups sum of squared errors (WGSS).

Minimize $P(W,Q)=\sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l}d(X_i,Q_l)$

Subject to $\sum_{l=1}^{k} w_{i,l} \quad 1 \le i \le n$

$w_{i,l} \in \{0,1\}, \quad 1 \le i \le n, 1 \le l \le k$

6. Experimental Results

6.1. Classification Performance

The dataset are stored in the database with 10 fields and data relevant to that field. The age is very important to identify the diabetics for the person. The data set is containing ten attributes namely name, preg, plas, mass, age, insu, skin, pedi, pres and classlab.

The data set is classified using the algorithm and attain the result may tested_positive or tested_negative. This result is compare with the original classlab of that specific data set, if both are matches we classify the exactly, then its count as True Positive.

Likewise count the values for True Negative (TN), False Positive (FP) and False Negative (FN).Then calculate the classification rate using this formula:

- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- Measure = 2*TP/ (2*TP+FP + FN)

RECALL is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage. PRECISION is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.

True Positive means that the data is exactly classified. False Positive means that an unexpected result is achieved after classification done. False Negative means that the missing value of the classification. It means some of the

values cannot be classified. True Negative means that the correct classification of the absence of result.

I took 20 samples to test this algorithm, it exactly classify the all the samples. This algorithm need to classify the data set has 768 instances, each being described by 10 attributes. The instances were classified into two classes, approved labeled as "tested_negative" and "tested_positive".

6.2. Clustering Performance

The primary use of clustering algorithms is to discover the grouping structures inherent in data. The advantage of this approach is the structures of constructed data sets can be controlled.

	Cluster 1	Cluster 2
Tested_negative	373	127
Tested_positive	118	150

This table is obtained using WEKA tool. It also clusters the data set according to this result.

The 768 sample data set and its clustered into 3 cluster using the distance measure. Before clustering the pre processing is done using normalization method.

In this algorithm using distance measure, the dataset are clustering into three groups. Initialize the cluster values at randomly and cluster the remaining values using distance formula.

In WEKA 7.6 tool, classified this data set at 70 % of classification rate. So now it's classified using some criteria which is used to increase the classification rate.

$$\sum_{i=1}^{n} d1(Xi, Q) = \sum_{i=1}^{n} \sum_{j=1}^{m} \delta(x_{i,j}, q_j)$$

6.3. Output

This shows the output for sample 30 diabetic's dataset. The data are saved in the Ms-Access. The database is connected through JDBC and retrieved. The data are processed and calculate the classification rate.

D:\JAVA\JDK1.3\bin>javac classificationrate.java

D:\JAVA\JDK1.3\bin>java classificationrate

Attempting to load JDBC Driver....

JDBC Driver loaded...

Connecting to database...

Database connection established

i=30

Connection to DB closed. Data Retrieved Successfully!

Classification Result*******

tested_negative	ani
tested_negative	devi
tested_negative	kavi
tested_positive	pavi
tested_negative	mani
tested_negative	vimal
tested_negative	ravi
tested_negative	kumar
tested_positive	jansi
tested_negative	janaki
tested_positive	santhiya
tested_positive	sudar
tested_negative	gukan
tested_negative	kanika
tested_negative	dhivya
tested_negative	murali
tested_negative	sankar
tested_negative	yuvi

The classification rate is: 66.66666666666666%

****Clustering Result**********

Data is classified into 3 clusters as follows.

Cluster 1

---

preg plas

6.0 148.0

0.0 137.0

5.0 116.0

10.0 115.0

8.0 125.0

4.0 110.0

10.0 139.0

0.0 118.0

1.0 115.0

Cluster 2

---

preg plas

1.0 85.0

1.0 89.0

3.0 78.0

7.0 100.0

7.0 107.0

1.0 103.0

Cluster 3

---

preg plas

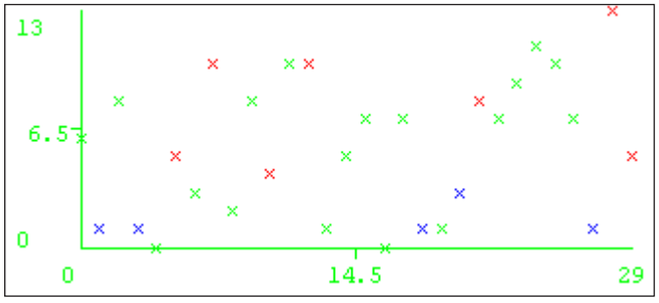
8.0 183.0

2.0 197.0

10.0 168.0

1.0 189.0

5.0 166.0



Clustering result for 30 samples

7. Conclusions

The most attractive property of the *k*-means algorithm in data mining is its efficiency in clustering large data sets. Classification is a data mining technique used to predict group membership for data instances. The classification is done using this algorithm and successfully classified the data set into two class labels namely tested_positive and tested_negative.

The clustering performance of the two algorithms has been evaluated using two real world data sets. The satisfactory results have demonstrated the effectiveness of the algorithms in discovering structures in data

This paper has focused on the technical issues of extending the *k*-means algorithm to cluster the diabetic's data set and classify the dataset. After that, using this algorithm calculate the classification rate. For the 20 sample data set it gives 100% classification rate. For the whole data set it must be improved and reached that level.

The proposed algorithm, is used to improved the classification rate and achieve the 100% result. Also efficiently cluster the dataset using k-modes algorithm and combined k-means and k-modes algorithm. It mainly help to improve the efficiency of the clustering the dataset.

References

1. Huang, Z. (1998). Extensions to the *k*-Means Algorithm for Clustering Large Data Sets with Categorical Values, *Data Mining and Knowledge Discovery*, 2, 283-304.
2. Mitchell, T. (1997). *Decision Tree Learning* (52-78). McGraw-Hill Companies, Inc.
3. Yasodha, P. & Kannan, M. (2011). *Analysis of a population of diabetic patients databases in Weka tool.* Proceedings of the International Journal of Scientific & Engineering Research, 2(5).
4. *Editorial, Diagnosis and Classification of Diabetes Mellitus, American Diabetes Association, Diabetes Care*. (2004). 27(1).
5. Karegowda, A. G., Punya, V., Manjunath, A. S. & Jayaram, M. A. (2012). Rule based classification for diabetic patients using cascaded K-means and decision tree C4.5. *International Journal of Computer Applications*, 45(12), (0975-8887).
6. Karegowda, A. G., Jayaram, M. A. & Manjunath, A. S. (2012). Cascading K-means clustering and K-nearest neighbor classifier for categorization of di-

abetic patients. *International Journal of Engineering and Advanced Technology,* 1(3).

7. Wu, C., Steinbauer, J. R. & Kuo, G. M. (2005). *EM Clustering Analysis of Diabetes Patients Basic Diagnosis Index.* Articles from AMIA Annual Symposium Proceedings are provided here courtesy of American Medical Informatics Association.
8. Maseri, W., Mohd, W., Herawan, T. & Ahmad, N. (2013). *Applying Variable Precision Rough Set for Clustering Diabetics Dataset.* In: AST2013 and Soft-tech 2013 International Conference.
9. Vijayalakshmi, D. & Thilagavathi, K. (2012). *An Approach for Prediction of Diabetic Disease by Using b-Colouring Technique in Clustering Analysis*. Proceedings of International Journal of Applied Mathematical Research, 1(4), 520-530.