K-mean Clustering for Data Mining: A Review

[1]Er. Shilpy (Author), [2]Er Rashim Rana
[1,2]Department of Computer Science and Engineering
Maharishi Ved Vyas Engineering College – Jagadhri

Abstract: **Medical data mining is significant research area, there is a need to mine the medical data to extract useful patterns for disease prediction. Generally, numbers of tests are necessary from a patient for detecting a disease. These numbers of tests can be reduced by using data mining to improve the time and performance. It takes less time in predicting a disease like heart disease, diabetes, breast cancer etc. In this paper various data mining techniques E.g. K-means, Support vector machines, Genetic Algorithm etc. are discussed for predicting the disease. The various challenges faced by the disease diagnosis using data mining and the existing models of data mining used for prediction are also stated in this paper.**

Keywords: KDD, Prediction analysis, K-means

I. INTRODUCTION

There is very large amount of data everywhere around us. This data is coming from various sources like medical field, image processing, social websites etc. Data mining in the field of medical diagnosis is a major research area. It is a process of discovering interesting patterns from this large amount of data. It is extraction of implicit unknown and useful information from data. Data mining is also called as extraction of hidden patterns. It is also known as knowledge mining, knowledge extraction and data/pattern analysis. This process is fully automated or semi-automated to discover knowledge that is useful for user [2].

It is also a process of finding hidden information from the data base; it may use one or more computer learning techniques to automatically analyze and extract knowledge from the data in the database and it is a part of knowledge discovery process. In data mining algorithms are applied to large amount of data so as to produce models or patterns that are interesting to the user and will extract the hidden patterns.

1.1 Data mining KDD

Knowledge discovery process is an iterative process which contains following steps.

- Data cleaning: - In this step, noisy and inconsistent data is removed for large data, fill in the missing values and identify the outliers by using various techniques.
- Data integration: - In this, data from multiple data sources are combined. It is process of merging data from various data sources.
- Data selection: - In this step, data that is relevant and required by the analysis task are retrieved from the database.
- Data transformation: - In this step, data are transformed and consolidated into forms that is required or appropriate for mining by performing various operations like normalization, aggregation, generalization and attribute construction
- Data mining: - It is very essential process in which hidden and useful patterns are extracted from large databases for representing knowledge.
- Pattern evaluation: - In this, knowledge is represented by using interesting pattern based on various measures.

1.2 Need of data mining

- Very large data amount of data are generated every day. Data Mining is used extract useful information and knowledge from vast amount of data
- Data that are generated have different dimensionality. To deal with these types of dimensionality Data Mining is used.
- Variety of data is generated every day. So to deal with heterogeneity of data. Data Mining is used.

1.3 Data mining process

An iterative process which include the following steps

- State the problem for example:-classification or numeric calculation
- Assemble the data relevant that relate with stated problem.
- Perform preprocessing on data and represent the data in the form of labels.

- Study a model or predictor.
- Assess the model.
- Well adjust the model as needed

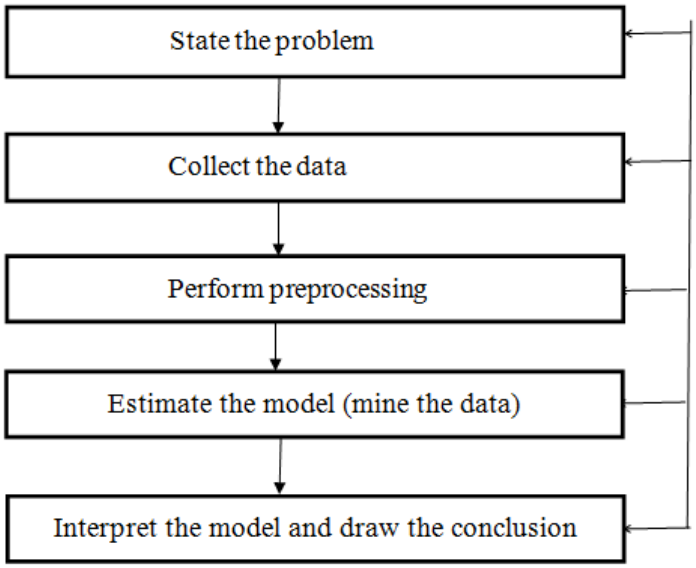

Figure 1.1 Data mining process

1.4 Clustering in Data Mining: Clustering is an unsupervised learning technique. It is a process of partitioning a set of data into a set of significant sub-classes, called cluster. Data is organized into clusters such that there is high intra-cluster similarity and low inter-cluster similarity [9]. It is implemented in many fields including machine learning, pattern recognition, image analysis, information retrieval and bioinformatics.

K-mean clustering algorithm: K-means clustering algorithm is one of the most popular and simplest algorithms. It is unsupervised learning algorithm that is used to solve the sound known clustering problems. Procedure followed by it a very simple and easy way to classify a given data set. K-mean clustering algorithm has some properties that are specified below:

- There should be always k cluster.
- Each cluster always contains at least one item.
- Non-hierarchical clusters are formed and they do not overlie

Algorithm: K-means:-The k-means algorithm is one of partitioning algorithm, in which each cluster's center is represented by the mean value of the objects in the cluster [1].

Input:

K: represent number of clusters,

D: specify a data set contain n objects.

Output:

A set of k clusters are generated. Method:

- Choose k data objects representing the cluster centroids.
- Assign each data object of the entire data set to the cluster having the closest centroid.
- Compute new centroid for each cluster, by averaging the data objects belonging to the cluster.
- If at least one of the centroids has changed, go to step 2, otherwise go to step 5
- Output the clusters.

Advantages of Kmeans:

- Simple and easier to understand.
- Fast and robust.
- Relatively efficient and gives best results if the data sets are distinct from each other.

Challenges in K-means:

- K-means algorithm assumes that the number of clusters k in the database is known beforehand which, obviously, is not necessarily true in real-world applications

- As an iterative technique, the k-means algorithm is especially sensitive to initial centers selection
- K-means algorithm may converge to local minima.

1.5 Prediction analysis: It includes a variety of techniques that analyze the historical data and the present data to make predictions. It can be applied to any unknown event whether it is in past, present or future. For prediction, the first step is to build a predictive model or classifier that can best predict the outcome. To get better results of classifier, there is a need of reduced feature set that limits the number of input features. This is termed as feature selection or feature extraction.

Prediction involves two steps:

- Building a model or classier: Model is built from the training set that contains database tuples and each tuple is known as class, category or data points.
- Using the model or classifier: Test set is used to find the accuracy of the classification rules. If the rules are acceptable these can be applied on new data [10].

II. DISEASE PREDICTION

The predictive data mining for medical diagnosis is a significant research area. There is a very important role of data mining for predicting diseases. Generally, numbers of tests are necessary from a patient for detecting a disease. These numbers of tests can be reduced by using data mining to improve the time and performance. It takes less time in predicting a disease like heart diseases, diabetes, breast cancer etc. using data mining techniques. The classifier systems e.g. Support Vector Machine are highly used in medical diagnosis. These help in minimizing the errors and also examine the medical data in shorter time. The methods of medical data mining include genetic algorithms, artificial neural networks, fuzzy systems and support vector machines. The missing value imputation method such as case deletion, most common method, k-means clustering imputation, k-nearest neighbor etc. can be applied on dataset to fill the missing data values in the data set before pattern extraction and prediction. Missing value can occur because of many reasons e.g. error in manual data entry, incorrect measurements or equipment errors and can produce problems in the data mining process. The two important algorithms used in disease prediction are genetic algorithm and support vector machine. Genetic Algorithm (GA) is used for selecting the features of interest in predicting the value. It works on the principal of genetics and can have multiple offspring. While generating the solutions, it follows "Survival of Fittest" principal. Solutions are either 0 or 1. Genetic algorithm requires fitness function for evaluating solution domain. Support Vector Machine (SVM) is a statistical learning method that classifies the cases of different class labels by constructing hyperplanes. The subset of data instances that is used to define the hyperplane is known as, support vector and the distance between the nearest support vector and the hyperplane is called margin.

2.1 Challenges in disease prediction using data mining [10]:

In disease prediction, poor decisions can cause disastrous consequences so there is a need of high degree of accuracy. Following are some of the challenges in prediction through data mining:

- The medical data is gathered from various sources e.g. conversation with the patients, laboratory results and interpretation of doctors. All these have major impact on treatment of the patient. Decisions based on these may fail in some cases.
- Selecting the reduced set of features or attributes before classification on machine learning algorithm is a major challenging task.
- Large data set having missing values is also another challenge to be conquered in terms of computation time.
- To provide high quality of service. Quality of service implies diagnosing disease correctly & provides effective treatments to patients.

2.2 Existing disease prediction models

Model	Description	Accuracy (%)
SVM	Machine learning method that classifies disease from the high dimensional medical data.	78
GA+SVM	GA is used for feature selection and SVM is used for classification	77.3
K-means+ SVM	K-means is used for feature selection and SVM is used for classification	93.65

Modified K-means+ SVM	Modified K-means is used for feature selection and SVM is used for classification	96.71
K-means+ GA+SVM	K-means is used to remove outliers and noisy data, GA is used for feature selection and SVM is used for classification	98.82

III. RELATED WORK

Zheng B, Yoon S. W. and Lam S. S. in 2014 purposed [1] a hybrid of K-means algorithm and support vector machine algorithm for feature extraction. K-means is used for recognizing the hidden patterns of tumor. The tumor feature data set is classified into malignant (cancerous) and benign (aren't cancerous) sets separately. The membership function is used to find the similarity between incoming tumor and symbolic tumor, and obtains the compact result of k-means. Support Vector Machine (machine learning algorithm) is now applied on reduced feature space. This purposed K-SVM model gives higher accuracy with reduction in computation time.

T. Santhanam, M. S. Padmavathi in 2015 implemented [2] K-means along with Genetic Algorithm for dimensionality reduction and support vector machine to classify the data set. K-means algorithm is used to remove outliers and the noisy data. The optimal features are selected by using the genetic algorithm and then Support vector machine classifies the reduced data space using 10 fold cross validation technique. Genetic algorithm selects different features from original set of feature during each run. To obtain consistent results, the experiment was performed 50 times. The result shows that the purposed model achieves the accuracy of 98.82%.

A. Purwar, S.K.Singh in 2015 proposed [4] a prediction model for medical data with missing value imputation techniques, then analyzing these techniques by using K-means algorithm and choosing the best among them. Thus this model improves the quality of data by using the best imputation technique. Methods such as case deletion, most common method, concept most common, K-means clustering imputation, k-nearest neighbor etc. are applied to fill the missing data values in the data. The efficiency is calculated on three data sets namely Hepatitis, Wisconsin Breast Cancer and Pima Indians Diabetes from the UCI repository. This model achieved accuracy of 99.82% for Diabetes data set, 99.39% for Breast Cancer and 99.08% for Hepatitis data set. For Diabetes and Hepatitis data sets Concept Most Common (CMC) is chosen as the best method, and for Breast Cancer Case deletion is selected as best missing value imputation method.

A K Yadav, D Tomar, S Agarwal [5] in 2013 diagnosis of lung cancer. The lung cancer dataset is discussed with the domain experts and certain attributes with their impact factor are identified based on which the number of cluster is decided e.g. there is a possibility of cancer if the tumor size is greater than 3. On the basis on this clusters are formed. Then the cluster will move left or right according to the impact of the next attribute on the cluster. The results show that the Foggy K-mean gives better result than the simple k-mean algorithm.

M.F.Akay in 2009 proposed [3] a system for breast cancer diagnosis by using support vector machine combined with the feature selection. Wisconsin breast cancer dataset from UCI repository is used for the experiment. This dataset contains nine features which are represented as a value between1 and 10. F-score for each feature is calculated and then sort the scores for the features. More discriminative feature has the larger F-score. Feature section is helpful in reducing the number of input feature in SVM classifier.

A Jain, A Rajavat, R Bhartiya, in 2012 proposed [6] modified k-mean clustering algorithm to cluster large datasets, the main motive is to find out the cluster centers which are very close to the final result for each iterative step. Modified k-mean clustering algorithm reduces problem of cluster error criterion and also avoids getting into locally optimal solution in some degree. They compare modified k-mean algorithm with k-mean clustering algorithm and the results shows that modified k-mean clustering algorithm take less time to execute than existing k-mean for small number of records as well as for large number of records. Modified k-mean algorithm is more strong to noise and outliers than K-means.

Poteras, C. M., Mihaescu, M. C., & Mocanu, M in 2014 proposed [7] an optimized version of k-mean that reduces the problem of re-distribution of the data elements that will remain part of the same cluster during the next iteration. After a number of iterations only a few number of data elements change their cluster. While assigning the data element to the cluster there is no need to visit the entire data set, but just a small list of data objects. The implementation showed up to 70% reduction of the running time.

S Bharti, S. N Singh in 2015 stated [8] several algorithms like genetic algorithm, PSO, ANN that can be used in predicting heart disease. Combining these algorithms with the data mining techniques such as clustering, classification etc. or by combining these algorithms with one another will give better performance and accuracy.

IV. CONCLUSION

Data mining plays a significant role in disease prediction from time and performance perspective. Large data set having missing values is a major challenge because prediction requires high degree of accuracy. The missing value imputation method such as case deletion, most common method, k-means clustering imputation, k-nearest neighbor etc. can be applied on dataset to fill the missing data values. K means is the simplest and the efficient clustering algorithm that can be used to cluster the data and removing outliers, inconsistent data and noisy data. Modified K-means algorithm can also be used in place K- means to improve its efficiency. Genetic algorithm generates the reduced set of features called feature extraction. It is beneficial to limit the number of input features in a classifier in order to have a good predictive model.

REFERENCES

[1] Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), 1476-1482.

[2] Santhanam, T., & Padmavathi, M. S. (2015). Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis. Procedia Computer Science, 47, 76-83.

[3] Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications, 36(2), 3240-3247.

[4] Purwar, A., & Singh, S. K. (2015). Hybrid prediction model with missing value imputation for medical data. Expert Systems with Applications, 42(13), 5621-5631.

[5] Yadav, A. K., Tomar, D., & Agarwal, S. (2013, July). Clustering of lung cancer data using Foggy K-means. In Recent Trends in Information Technology (ICRTIT), 2013 International Conference on (pp. 13-18).

[6] Jain, A., Rajavat, A., & Bhartiya, R. (2012, November). Design, Analysis and Implementation of Modified K-Mean Algorithm for Large Data-Set to Increase Scalability and Efficiency.In Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on (pp. 627-631).

[7] Poteras, C. M., Mihaescu, M. C., & Mocanu, M. (2014, September). An optimized version of the K-Means clustering algorithm. In Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on (pp. 695-699).

[8] Bharti, S., & Singh, S. N. (2015, May). Analytical study of heart disease prediction comparing with different algorithms. In Computing, Communication & Automation (ICCCA), 2015 International Conference on (pp. 78-82).

[9] Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16, 645–678

[10] Milovic, B., & Milovic, M. (2012). Prediction and decision making in Health Care using Data Mining. Kuwait Chapter of the Arabian Journal of Business and Management Review, 1(12), 126.