International Journal of Engineering & Technology, 7 (4.5) (2018) 40-43

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Real Time Symptomatic Analysis for Efficient Disease Prediction and Recommendation Generation Using Multi Level Symptom Similarity Measure

[1] *Sathish Kumar.P.J, [2] Dr.R.Jagadeesh Kannan

[1] Research Scholar, BIHER- Bharath Institute of Higher Education & Research, Selaiyur, Chennai, Tamilnadu
[2] Professor, CSE, VIT-Vellore Institute of Technology Chennai Campus, Chennai, Tamilnadu
**Corresponding author E-mail: [1] E-mail:sathishjraman@gmail.com*

Abstract

The problem of high dimensional clustering and classification has been well studied in previous articles. Also, the recommendation generation towards the treatment based on input symptoms has been considered in this research part. Number of approaches has been discussed earlier in literature towards disease prediction and recommendation generation. Still, the efficient of such recommendation systems are not up to noticeable rate. To improve the performance, an efficient multi level symptom similarity based disease prediction and recommendation generation has been presented. The method reads the input data set, performs preprocessing to remove the noisy records. In the second stage, the method performs Class Level Feature Similarity Clustering. The classification of input symptom set has been performed using MLSS (Multi Level Symptom Similarity) measure estimated between different class of samples. According to the selected class, the method selects higher frequent medicine set as recommendation using drug success rate and frequency measures. The proposed method improves the performance of clustering, disease prediction with higher efficient medicine recommendation.

Keywords*:* High Dimensional Clustering, Map Reduce, Disease Prediction, Symptoms, Recommendation, MLSS.

1. Introduction

The modern society suffers with various diseases which has been increasing in every day in numbers. However, the symptoms of different diseases are more similar in most cases. For example, the symptoms of generic fever like "Temperature", "body Pain" are more common with other diseases like "Typhoid" and "Dengue". This introduces challenges to the medical practitioner in providing treatment and performing medicine selection. However, there exist many number of drugs available for the same disease, the selection becomes more difficult for the practitioner.

In general case, the medical records of various patients would help in identifying the disease being affected by the patients. To perform disease prediction and to generate recommendations, it is necessary to cluster them. As the dimension of the medical records are higher, the map reduce is more necessary. There are number of map reduce algorithms available in literature. The most algorithms consider only few number of features in clustering the data points of medical data set. This would reduce the performance of clustering and affect the performance of recommendation system.

High dimensional clustering is the process of grouping the data points of medical data set. To perform clustering or map reduce, it is necessary to consider the maximum number of features in estimating similarity between the data points of cluster and input sample. To improve the performance in clustering and other issues, this paper perform clustering according to class level feature similarity measure which has been measured based on the similarity in each dimension of data points. Similarly for the disease prediction, an efficient multi level symptom similarity measure has been presented. The measure would be used in disease prediction as well as increase the performance in recommendation system.

The disease prediction is the process of predicting the possible disease would occur or based on the symptoms, identifying the possible disease. It can be performed in many ways when the dimension of the data point is less. When the number of symptoms is higher or the dimension of the data point is higher it will be difficult. Also, there will be similar symptoms for different diseases this increases the challenge in identifying the possible disease. This increases the requirement of strategically approach in disease prediction. The disease prediction algorithm has to consider all the features or symptoms available to identify the possible disease the patient have been affected by.

The recommendation system would be efficient when it uses the similarity measure in efficient manner. For the recommendation generation, the method has to count the success rate of different medicines in curing the disease. To achieve this, the success frequency measure based approach is presented in this paper. The detailed approach is discussed in the next section.

2. Related Works

There are number of approaches have been discussed for the recommendation system, and this section presents different approaches towards the issue considered.

Application of data mining methods in diabetes prediction [1], discuss various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. The methods strongly based on the data mining techniques can be effectively applied for high blood pressure risk prediction. In this paper, we

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

explore the early prediction of diabetes via five different data mining methods including: GMM, SVM, Logistic regression, ELM, ANN.

Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus [2], explores the early prediction of diabetes using various data mining techniques. The dataset has taken 768 instances from PIMA Indian Dataset to determine the accuracy of the data mining techniques in prediction. The analysis proves that Modified J48 Classifier provide the highest accuracy than other techniques.

Glucose prediction data analytics for diabetic patients monitoring [3], present a comprehensive critical review focusing on recent glucose prediction models and a best fit model is proposed based on the evaluation to perform data analytics in a wireless body area network system. The proposed glucose prediction algorithm is based on autoregressive (ARX) model which consider exogenous inputs such as CGM data, blood pressure (BP), total cholesterol (TC), low-density lipoprotein cholesterol (LDL), high density lipoproteins (HDL). A dataset of 442 diabetic patients is used to evaluate the performance of the algorithm through mean absolute error (MAE), root-mean-square error (RMSE), and coefficient of determination (R2).

Neuro-Fuzzy based Glucose Prediction Model for Patients with Type 1 Diabetes Mellitus [4], presents the design, the development and the evaluation of a personalized glucose prediction model for patients with Type 1 Diabetes Mellitus (T1DM). The personalized model is based on neuro-fuzzy techniques in order to capture the metabolic behavior of a patient with T1DM. Moreover, wavelets are applied as activation functions in order to enhance the prediction performance and avoid local minimum during training stage. The model receives as input, data from sensors which record in real time glucose levels and physical activity, and provides with future glucose levels.

Predictions in heart disease using techniques of data mining [5], intends to give details about various techniques of knowledge abstraction by using data mining methods that are being used in today's research for prediction of heart disease. In this paper, data mining methods namely, Naive Bayes, Neural network, Decision tree algorithm are analyzed on medical data sets using algorithms.

Predicting disease by using data mining based on healthcare information system [6], applies the data mining process to predict hypertension from patient medical records with eight other diseases. A sample with the size of 9862 cases has been studied. The sample was extracted from a real world Healthcare Information System database containing 309383 medical records. We observed that the distribution of patient diseases in the medical database is imbalanced. Under-sampling technique has been applied to generate training data sets, and data mining tool Weka has been used to generate the Naive Bayesian and J-48 classifiers. In addition, an ensemble of five J-48 classifiers was created trying to improve the prediction performance, and rough set tools were used to reduce the ensemble based on the idea of second-order approximation.

Prediction of sugarcane diseases using data mining techniques [7], presents a succinct study of sugarcane disease calculation by Decision Tree Model (DTM) method and Random Forest method for India. These Data mining techniques give us a better solution to this problem, which can be applied to increase the sugarcane yield prediction.

Decision Support System for Heart Disease Prediction using ata Mining Techniques [8], discuss that Disease finding is one of the applications where Data Mining devices are demonstrating successful results. These are the main reason for death everywhere throughout the world in the past ten years. Several scientists are utilizing factual and Data Mining apparatuses to over assistance social insurance experts in the analysis of this disease. Using Hybrid Data Mining strategy in the analysis of coronary illness has been completely explored indicating satisfactory levels of accuracy.

Heart Disease Prediction System using Data Mining Method [9], proposes a HDPS based on three different data mining techniques. The various data mining methods used are Naive Bayes, Decision tree (J48), Random Forest and WEKA API. The system can predict the likelihood of patients getting a heart disease by using medical profiles such as age, sex, blood pressure, cholesterol and blood sugar. Also, the performance will be compared by calculation of confusion matrix.

In [10], the author developed a Decision Support in Heart Disease Prediction System (HDPS) using data mining modeling technique, namely, Naïve Bayes. Using medical profiles such as age, sex, blood pressure and blood sugar, chest pain, ECG graph etc it can predict the likelihood of patients getting a heart disease. It is implemented in matlab as an application which takes medical test's parameter as an input. It can be used as a training tool to train nurses and medical students to diagnose patients with heart disease.

3. MLSS Based Disease Prediction and Recommendation

The proposed MLSS based approach groups the medical data set into number of disease class according to class level feature similarity measure. In the second stage, the method takes the input symptoms and estimates multi level symptom similarity measure to identify the disease class. Finally, the method computes the drug success rate to select medicines and generate recommendations to the medical practitioner.

3.1. Preprocessing

In this stage, the input data set has been read and the list of attributes of the data points is identified. Then, for each data point, the method verifies the fulfill on all the dimensions. If any of the data point is identified as incomplete and noisy, it will be removed from the dataset. The preprocessed data set will be used to perform clustering and disease prediction.

3.2. CLFS Clustering

The class level feature similarity based clustering algorithm, reads the input data set. For each data point, the method estimates, multi level class feature similarity (MLCFS) on all the dimensions. Based on the MLCFS estimated, the method identifies the class of the data point and assigns the data point to the class identified. The generated cluster has been used to perform disease prediction.

3.3 Algorithm

Input: Preprocessed data set P, Cluster C
Output: Null.
Start
Read P, Initialize C.
For each data point p
For each disease class d of C

$$\text{ComputeMLCFS} = \frac{\sum_{i=1}^{size(D)} \sum_{j=1}^{size(p)} Dist\left(p(j), D(i)(j)\right) < Threshold \,/\, size(p)}{size(C)}$$

End
Choose the class C with higher MLCFS.
Assign data point to class C.

$$C = \sum (\text{Datapoint} \in C) \cup p$$

End
Stop

The above discussed algorithm computes the multilevel class feature similarity measure on different class data points. Finally a single one is selected based on that and the data point is indexed to the identified class.

4. MLSS Disease Prediction and Recommendation

The disease prediction algorithm reads a set of symptoms and their values. Based on the symptoms , the method computes multi level symptom similarity measure with the disease class set available. Based on the MLSS measure, a single disease class has been selected. From the disease class identified and the list of medicines available, for each of them the method compute medicine success rate and frequency. Based on the success rate, a subset of medicines has been recommended for the practitioner.

4.1 Algorithm

Input: Cluster set Cs, Symptoms S
Output: Recommendation R.
Start
Read Cluster set Cs.
For each disease class D

Compute MLSS=

$$\frac{\sum_{i=1}^{size(D)} \sum Symptoms(D(i)) == \sum Symptoms(S) \Big/ size(S)}{Size(D)}$$

End
Class c = Choose the disease class with maximum MLSS.
Find the list of medicines ML =

$$\sum_{i=1}^{size(Medicine\ Track)} Medicine(c) \in MedicineTrack$$

For each medicine m

Compute frequency fr = $\frac{\sum_{i=1}^{size(MedicineTrack)} MedicineTrack(i)==m}{Size(MedicineTrack)}$

Compute success rate Sr.

Sr = $\frac{\sum_{i=1}^{size(MedicineTrack)} MedicineTrack(i)==m \&\& Impact==Success}{Fr}$

End
Recommendation= Sort the medicines based on success rate and populate.
Stop.

The proposed algorithm computes the success rate for different medicines and based on that the medicines are sorted to the practitioner

5. Results and Discussion

The proposed MLSS based disease prediction algorithm has been implemented and evaluated for its efficiency using different test cases. The method has produced efficient results in disease prediction and classification accuracy. The result produced has been presented below and the test conditions are listed in the table 1.

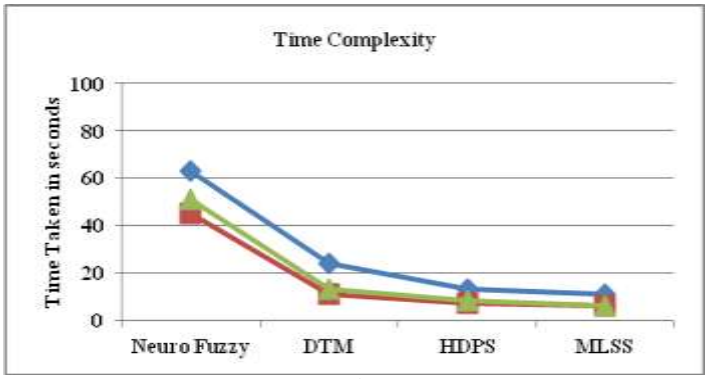
Table 1: Description of Data set used

Parameter	Value
Data Set Name	UCI
No of Dimensions	23174
No of Data Points	275
No of classes	5

The Table 1 describes the information about the data sets used for the evaluation of the proposed method. The data sets have varying number of data points and each has different number of data points and number of gene numbers. With the UCI data set, we have collected the lifestyle and physical feature. Also, the conditions of eye and pancreas have been collected to collaborate with the UCI data set. Using the cooked data set, the performance of the algorithm has been measured.

5.1 Time Complexity

The time complexity of classification and prediction has been measured. It has been estimated based on the time value taken for prediction with X number of samples. It has been approximated for different number of samples.

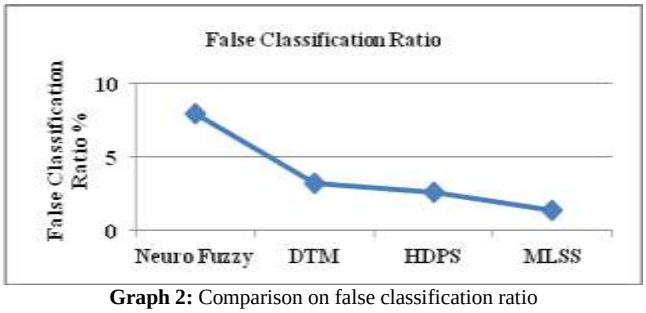


Graph 1: Comparison of time complexity

The Graph 1, shows the comparative result on time complexity produced by various methods and the result shows that the proposed method has produced less time than other methods.

5.2 Classification Ratio

As the method includes the higher number of features, the false classification ratio will get reduced. The false classification ratio will be reduced and the same implies on the increase of prediction accuracy.

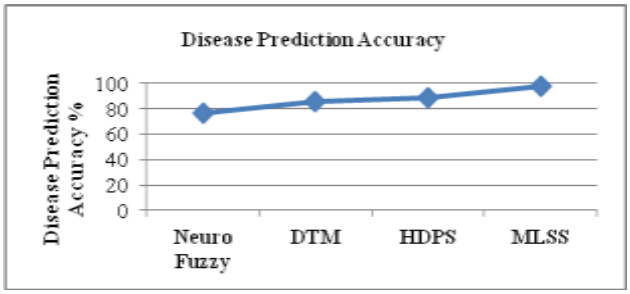


Graph 2: Comparison on false classification ratio

The Graph 2, shows the comparative result on false classification ratio produced by various methods and its shows clearly that the proposed method has produced less false classification ratio than other methods.

5.3 Prediction Accuracy

The prediction accuracy has been increased, by including different characteristic features. As the algorithm includes all the symptoms in measuring the MLSS value, the prediction accuracy will be increased.

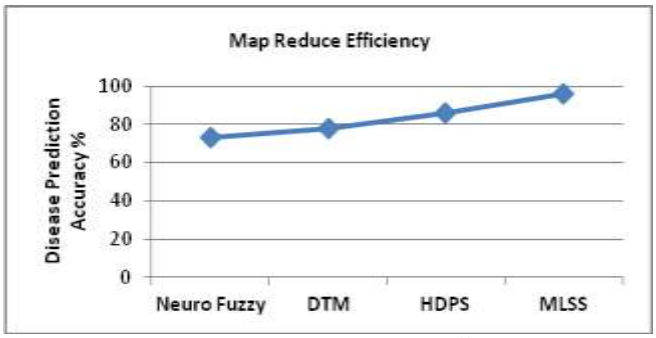


Graph 3: Comparison on disease prediction accuracy

The Graph 3, shows the comparison on disease prediction being produced by different methods. The proposed method produces higher disease prediction accuracy than other methods.

5.4 Maps Reduce Efficiency

The map reduce efficiency is improved because it uses the CLFS measure. In the earlier system it is performed based on the known classes.



Graph 4: Comparison on map reduce efficiency

The Figure 4, shows the comparative result on map reduce produced by different methods. The proposed MLSS algorithm has produced higher map reduce efficiency than other methods.

6. Conclusion

In this paper an efficient class level feature similarity and multi level symptom similarity based disease prediction and recommendation algorithm is presented. The method cluster the data points of the data set based on CLFS measure. For the classification and disease prediction, the method computes the multi level symptom similarity measure on all the disease class. Finally a single disease class is selected. From the disease class, the method computes the success rate for different drugs and populate according to them to the practitioner. The method improves the performance of mapreduce, disease prediction and recommendation.

References

[1] MessanKomi;JunLi;YongxinZhai ; XianguoZhang,Application of data mining methods in diabetes prediction, IEEE Conference on Image, Vision and Computing (ICIVC), 2017 .

[2] Devi, M. Renuka, J. Maria Shyla, "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus", International Journal of Applied Engineering Research, vol. 11, no. 1, pp. 727-730, 2016.

[3] GeshwareeHuzooree;KaviKumarKhedo;NoorjehanJoonas ,Glucose prediction data analytics for diabetic patients monitoring, IEEE Conference on Next Generation Computing Applications (NextComp), 2017.

[4] K. Zarkogianni, K. Mitsis, A. Fioravanti, K. S. Nikita, "Neuro-Fuzzy based Glucose Prediction Model for Patients with Type 1 Diabetes Mellitus", Ieee, pp. 252-255, 2014.

[5] Monika Gandhi ; Shailendra Narayan SinghPredictions in heart disease using techniques of data mining, IEEE Conference on : Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015.

[6] Feixiang Huang ;Shengyong Wang ;Chien-Chung Chan,Predicting disease by using data mining based on healthcare information system, IEEE Conference on Granular Computing, 2012.

[7] R. Beulah ; M. Punithavalli,Prediction of sugarcane diseases using data mining techniques, IEEE Conference on Advances in Computer Applications (ICACA), 2017.

[8] AnkurMakwana and Jaymin Patel. Decision Support System for Heart Disease Prediction using Data Mining Techniques. International Journal of Computer Applications 117(22):1-5, May 2015.

[9] Keerthana T K "Heart Disease Prediction System using Data Mining Method", International Journal of Engineering Trends and Technology (IJETT), V47(6),361-363 May 2017.

[10] Ms.RupaliR.Patil "Heart Disease Prediction System using Naïve Bayes and Jelinek-mercer smoothing" IJARCCE 2014.