Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/130511

Efficient Classification of Medical Data and Disease Prediction Using Multi Attribute Disease Probability Measure

K. Ananthajothi[1,*] *and M. Subramaniam*[2]

[1] Department of Computer Science and Engineering, Misrimal Navajee Munoth Jain Engineering College,Chennai, India.
[2] S.A.Engineering College, Chennai, India.

Received: 23 Jan. 2019, Revised: 31 May 2019, Accepted: 13 Jun. 2019
Published online: 1 Sep. 2019

Abstract: In this paper, we introduce a multi attribute disease probability -based classification approach. The method first identifies the list of features available in the data set. Based on identified features, the data points have been verified for their completeness with all the features identified. The data points having incomplete and missing features have been eliminated from the data set. Further the method computes the probability measure on each dimension. According to the probability measure of various dimensions, the multi attribute disease probability (MADP) has been measured for each disease class. It is shown that the proposed MADP -based disease prediction algorithm achieves higher classification ratio and disease prediction accuracy.

Keywords: Medical Data, High Dimensional Data, Classification, Disease Prediction, MADP.

1 Introduction

The modern world has great threat from various diseases which unknowingly affect the human health. The medical practitioner would not know or conclude everything in a single visit. This increases the requirement of automatic systems which would provide recommendations to the medical practitioner [1].The medical practitioner would submit the set of symptoms and their values where the system would produce a result. Based on the result produced, the medical practitioner would provide treatment to the patients [2]. The classification of the symptom depends on the values of the symptoms.

In general, the medical data covers various diagnosis results and symptoms. The size of medical data is higher and its volume increases at each movement. The medical organizations maintain vast amount of diagnosis results belonging to different people. Such huge data set can be used to perform disease prediction and classification. In order to perform classification, it must be measured for the similarity with the data points of any class. The similarity of data points towards the data points of a class has been measured in various methods. The basic nearest neighbor algorithm computes similarity using Euclidean distance measure which computes distance between the points. Similarly there is a number of similarity measures available to support the classification of data points.The deep similarity learning frameworks which simultaneously learn patient representations and measure pairwise similarity [3]. In [4], the data set considered has been clustered using K-means algorithm and identifying the correlated features to perform this is produced. However, K-means algorithm is not suitable to produce higher classification in the sense they do not measure the similarity efficiently, as the feature selection is not performed optimally.

Moreover, the previous similarity estimation algorithms consider only few dimensions of any data point. But in reality the medical data has higher number of dimensions which has no condition. In this case, measuring the similarity of data points in particular small set of dimensions would not help in achieving higher performance in classification or disease prediction. The blood glucose level has been predicted using personalized model to perform diabetic prediction [5]. However this does not consider the maximum possible features in prediction. It is necessary to consider almost all the dimensions in measuring the similarity of data points.

* Corresponding author e-mail: kanandjothime@gmail.com

This would support improvement of prediction performance. The possibility of disease being affected can be estimated based on the symptoms and their values. The accuracy and performance of prediction highly depends on the volume of symptoms considered.

This paper presents a probabilistic measure -based disease prediction and classification algorithm. The next sections will discuss the approach in detail.

2 Related Works

A number of methods has been available for the problem of medical data classification and disease prediction. In this section, a set of articles published related to the issue of disease prediction has been reviewed.

The influence of data mining in the problem of diabetic prediction has been analyzed. The methods considering the test results of chemical and blood have been considered. The data mining techniques have been applied over the higher blood pressure. Further, various data mining techniques have been used to predict diabetes in early stage [6].

The methods of diabetes mellitus prediction have been analyzed in [7], which uses PIMA data set with 768 instances to evaluate the performance of different data mining algorithms in diabetes mellitus prediction. The J48 classifier has been applied for classification and its performance has been measured.

The PCA, ICA and LDA, DWT methods are analyzed for their performance in the detection of CAD (Coronary Artery Disease) in [8]. The methods use the signals of heart rate which has been split into a number of sub bands using DWT. The features from sub bands have been feed to the classifier to measure the accuracy.

Curvelet transform -based facial expression recognition has been proposed, which uses spherical clustering. The curvelet transform has been adapted with RBF – Radial Basis Function to identify the hidden signals. The facial image has been split into different parts using curve let transform and spatial clustering has been used to perform expression recognition [9].

Towards monitoring the diabetic patients for their glucose prediction an efficient approach is presented in [10]. The author presents the detailed review on various glucose prediction model which uses regression analysis to predict diabetic with cholesterol and blood sugar.

A glucose monitoring algorithm is presented which considers the changing rate towards analyzing the risk. The markov chain algorithm has been used to visualize the variation and performs prediction based on the features considered [11].

Psychological features -based disease prediction algorithm is presented which considers the psychological features. The method considers the dynamic changes of blood glucose to perform classification with support vector regression model [12].

A comparative study on sensor -based monitoring model which uses sensor augmented pumps therapy. The signals received have been analyzed and used to generate alarms for the controller about various events [13]. Similarly a monitoring model for glucose concentration prediction in human has been presented. The method uses the continuous glucose monitoring with varying studies on both type 1 and 2 diabetes. The result of monitoring has been used to predict diabetes [14].

A glucose prediction model has been presented towards mobile platform. The popular support vector regression analysis is recommended for classification [15]. A comparative study on autoregressive models which use glucose and insulin inputs is presented. The neural network has been used to perform prediction which accounts both inter and intra variations [16].

The author presents a glucose prediction model based on neuro fuzzy technique for the prediction on Type 1 diabetes patients. The method uses the metabolic behaviors of the patients to generate personalized prediction using neuro fuzzy algorithm [17]. The inter and intra variation in glycemic values has been considered which are varying in time series. The integrated values have been used to perform prediction on glucose with grid analysis [18].

The least mean square algorithm -based learning approach is used for classification and disease prediction. The method uses different weights and different scientific approaches in classification [19]. Similarly a neural network -based diabetic detection system with mobile devices has been presented. The mobile devices attached with the human body sense the glucose level and transmit the data to the remote system which predicts the disease and based on that the glucose level of the patient has been controlled [20].

A multi-level approach for medical data classification is presented which uses influence measure in different level to identify the class of data points in an iterative manner [21].

Amarkov chained model has been presented to monitor the fluctuating glucose level with autoregressive with exogenous inputs (ARX) model. The parameters have been identified using expectation maximization algorithm [22].

Freestyle Libre (FSL) system is a new method to detect glucose enabling a new paradigm in glucose monitoring and self-management. The sensor, reading the interstitial fluid glucose concentration, provides a numerical data of glucose level and a trend arrow that add context to static measurement of glucose level. Therefore, patients could easily follow the progression of their glucose levels over the time, allowing early detection and timely treatment of deviations from targeted glucose level range, thus preventing extreme fluctuations [23].

All the above methods suffer to achieve higher performance in disease prediction and produces higher false classification ratio.

3 MADP Based Classification and Disease Prediction

The proposed multi attribute disease probability -based algorithm reads the input data set. From the input medical data set, the noises like incomplete data points are removed. Then, for each data point, the method computes disease probability with each dimension. Based on the computed disease probability values, the method estimates MADP (multi attribute disease probability) towards each class. The class which has higher probability has been indexed with the data point and identify the disease being affected by. The detailed approach is discussed in this section.

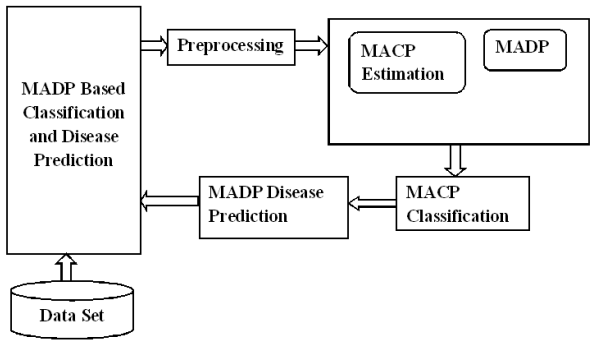

Fig. 1: Architecture of MADP classification and Disease Prediction.

Preprocessing:
In this stage, the method reads the input medical data set and identifies the list of values or dimensions present in the data set. The data points present in the data set have been verified such that it contains all the features that should be contained. If there exists any data point which misses some features then it will be neglected from the data set. Such data set without noise data point has been considered for the further classification and to be used for disease prediction.

Algorithm:
Given: Medical data set Md
Return: Preprocessed data set Prd

Start

Read data set Md.

Identify the list of dimensions or attributes Ads.

For each data point

$$Ads = \int_{i=1}^{size(Di)} \Sigma Dimensions(Ads) \cup \Sigma Dimensions(Di) \nexists Ads \quad (1)$$

End

For each data point Di

$$If \;\; Di \nexists \forall \, Ads \; then \quad (2)$$

Remove Di from Md

Else

$$Md = \Sigma(Di \in Md) \; \cup \; Di \quad (3)$$

End
End
Stop

The algorithm presented above presents how the preprocessing of input medical data set has been performed. The preprocessed data set becomes the input for further processing in disease prediction.

MADP Measure Estimation:
In this stage, for the given data point, towards the class given, the method estimates the probability. The method reads the data points of the class and with the input data point, and the method estimates the probability as follows: The value of the dimension in all the data points are considered and for each of them, the method estimates the distance measure. If the distance with the class value and input value are less than the distance threshold, then it has been counted. Based on the number of closure points and the total number of points of the class, the method estimates the probability measure.

Algorithm: Given: Class C, Data point Di
Return: MADP

Start

CDS = Read data points of class C.

Read data point Di.

Initialize count .

For each dimension Dim

Count =0;

For each data point of Dk from C

$$Computedimensionaldistance Dd = \sqrt{(Di.Dim - Dk.Dim)} \quad (4)$$

If Dd $>$ DTh then //Dth – dimensionality threshold

Count =count +1.
End
End

$$Computedimensionalprobability\ Dprob = count/(size(Dk)) \quad (5)$$

End

$$Compute\ \ MADP = \sum_{n=1}^{size(Di)} Dprob/size(Di) \quad (6)$$

Stop

The above method computes the probability measure of closure with all the dimension of data point. Based on the dimensional probability measure the MADP measure has been estimated.The disease probability on each dimension of feature has been computed based on the number of matches the data feature has with the samples of the data class. Estimated value has been used to perform classification and disease prediction.

MADP Disease Prediction and Classification :
In this stage, the method reads the input medical data set. The medical data set has been preprocessed to remove the noise from the data set. Then, for each disease class, the method estimates the MADP measure. Based on the MADP measure estimated, the method identifies a higher valued disease class to identify the disease. Estimated MADP measure has been used towards classifying the input data point.

MADP Disease Prediction/Classification Algorithm:
Given: Medical data set Md
Return: Disease class Dc

Start

Read medical data set Md

Pd = Preprocessing(Md)

For each disease class Dci

Compute multi attribute disease probability Madp.

End

Sc = Choose the disease class with higher MADP.

Index the data point to the class Sc.

Stop.

The above discussed algorithm estimates the multi attribute disease probability measure for each disease class available. By measuring the disease probability of a data point based on each feature considered, all the dimensions or features have been given importance and their representation in any disease has been considered.Based on the MADP measure a single class that has been selected and indexed.

4 Results and Discussion

The disease prediction algorithm with MADP has been implemented using matlab. The method has been validated for its performance in disease prediction using different data sets. Also, the method has been evaluated using varying number of test cases.

Table 1: Evaluation Details

Key	Value
Data Set Name	PIMA, UCI
Feature Dimension	9,14
Total Tuples	768,275
No of Disease classes	10

The proposed MADP prediction model has been evaluated for its performance in diabetic prediction. The evaluation has been performed based on the data set obtained from different scientific research units. The aim the data set has been released is to support diabetic prediction on various patients which contains many constraints. In PIMA data set all the patients are below 21 years and represent the diabetic. It has values of body mass index, insulin, age, number of pregnancy and so on. Similarly each data set UCI and PIMA has various constraints and they have been cooked to produce another data set which has been used perform disease prediction. Towards the evaluation of proposed and existing algorithm various factors has been considered. The data sets considered like PIMA, UCI has no common dimensions or attributes, so that they have been combined to produce the cooked up data set. They have been merged to produce a data set which has been used for evaluation.

Time Complexity:

The time complexity is the measure which represents the time taken for the classification of disease or prediction of the disease. The value of time complexity has been measured based on the time value for the classification of different samples.

Analysis on time complexity in disease prediction is performed with different methods. The result of analysis

Table 2: Analysis on time complexity

Method Name	PIMA	UCI
ARX	62	44
FSL	23	12
MLIIM	14	8
MADP	10	5

is presented in Table 2. The proposed MADP algorithm has achieved less time complexity compared to other methods.

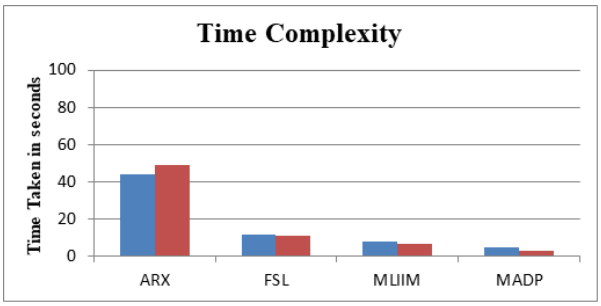

Fig. 2: Time Complexity Analysis.

Analysis on time complexity has been performed with different methods. The result of analysis is presented in Graph 1, which shows the proposed MADP algorithm has achieved less time complexity.The ARX, FSL, MLIIM and MADP algorithms have produced the time complexity in the ratio 62,23,14, 10 in PIMA data set and 44,12,8,5 towards UCI data set. In all the data set, the proposed MADP algorithm has produced less time complexity than any other method.

False Classification Ratio:

The false classification ratio is the measure which represents the frequency of incorrect classification. Because of the proposed MADP algorithm considered maximum number of features, the ratio of false classification gets reduced. This reflects on the performance of naïve bayes and svm. As the proposed algorithm used fuzzy rules, the ratio of false classification gets reduced. This in turn increases the prediction accuracy.

Table 3: Comparison on false classification ratio

Method Name	PIMA	UCI
ARX	7	9.3
FSL	4.1	6.2
MLIIM	2.7	3.2
MADP	1.4	1.5

The methods have been analyzed for their performance in false classification ratio. However, the proposed MADP algorithm has introduced less false ratio.The ARX, FSL,MLIIM and MADP algorithms have produced the false classification value in the ratio 7,4.1,2.7,1.4 towards PIMA dataset, 9.3,6.2,3.2,1.5 towards UCI data set. In all the cases, the proposed MADP algorithm has produced less false classification ratio than any other method compared.

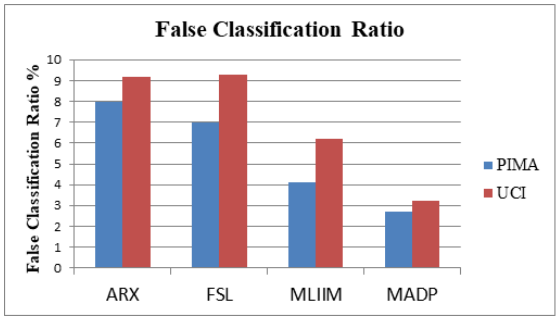

Fig. 3: Comparative Analysis on False Ratio.

Comparative analysis on false ratio has been measured with different methods. The result of analysis is presented in Graph 2. Finally, the proposed MADP algorithm introduces only negligible false ratio.

Prediction accuracy:

The accuracy on prediction has been estimated for different approaches according to the number of exact prediction and number of false prediction performed. As the method uses the fuzzy value considers more number of features, and the prediction accuracy has been increased. Also, the inclusion of lifestyle feature impacts on the prediction accuracy.

Table 4: Analysis on prediction accuracy

Method Name	PIMA	UCI
ARX	83	85
FSL	87	92
MLIIM	97.7	95
MADP	99.4	99.6

Analysis on disease prediction accuracy has been performed on different methods. The results have been presented in Table 3 and the MADP algorithm introduces higher prediction accuracy up to 99.5%. The proposed MADP algorithm considers various parameters in estimating the disease probability which has helped predicting the disease with more accuracy.The ARX,

FSL,MLIIM and MADP methods have produced the prediction accuracy in the ratio 82,88,97.6,99.6% towards UCI data set and 84,91,94,99.5% towards PIMA data set.

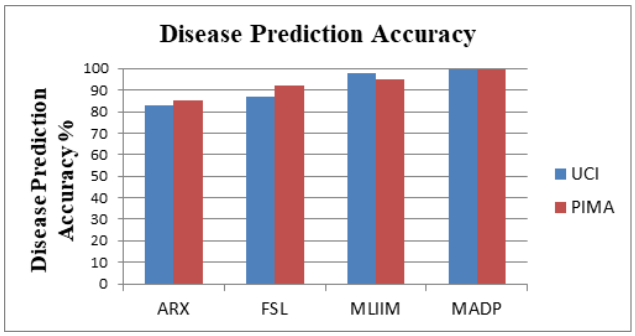

Fig. 4: Analysis on Disease Prediction Accuracy.

Analysis has been performed for the disease prediction accuracy of different methods. The inclusion of MADP in estimating the disease probability, the disease prediction accuracy has been improved. The result of analysis has been presented in Graph 3 and the proposed MADP algorithm has produced higher prediction accuracy.

5 Conclusion

In this paper an efficient MADP measure -based classification and disease prediction algorithm has been presented. The method performs preprocessing to remove the noise from the input data set. Then for each class data points, the method estimates the multi attribute disease probability measure. Based on the MADP measure being estimated, a single class has been selected with higher MADP measure as result. The method produces higher efficiency in map reduce, classification and prediction accuracy.As the method MADP considered multiple features in estimating the disease probability measure, the proposed MADP method has produced higher disease prediction accuracy up to 99.5% which in turn reduces the false classification ratio up to 0.5%. Similarly, the MADP algorithm can be used for the disease prediction on any disease.

ACKNOWLEDGMENT

This work was done in “Big Data and Cloud Computing Lab” at Dept. of Information Technology, S.A. Engineering college. The authors would like to thank, Department of Science & Technology, Ministry of Science & Technology, Govt. of India, for granting the fund under “Fund for Improvement of S&T Infrastructure in Universities and Higher Educational Institutions (FIST) Program – 2014”, Grant Sanction order vide: SR/FST/COLLEGE-239/2014, dated: 21st Nov 2014, for establishing “Big Data and Cloud Computing Lab” for strengthening the existing institutions S&T infrastructure and support for advancement in research works.

References

[1] Amir Kiani, Dynamic Recommendation: Disease Prediction and Prevention Using Recommender System, *International Journal of Basic Science and Medicine*, **1**, 1, 13-17, (2016).

[2] DhanashriGujar and RashmiBiyani, Disease Prediction and Doctor Recommendation System, *International Research Journal of Engineering and Technology (IRJET)*, **5**, 3, 3207-3209, (2018).

[3] QiulingSuo and Fenglong Ma, Deep Patient Similarity Learning for Personalized Healthcare, *IEEE Transactions on NanoBioscience*, **17**, 3, 219-227, (2018).

[4] Z. Fanmao, W. Y Ouqing, *Dynamic model with time varying delay for type 1 diabetes mellitus identified by using expectation maximization algorithm*, Presented at the 35th Chinese Control Conference, 9376-9381, (2016).

[5] Peter Gyuk, Istvan Vassanyi and Istvan Kosa, Blood Glucose Level Prediction for Diabetics Based on Nutrition and Insulin Administration Logs Using Personalized Mathematical Models, *Hindawi Journal of Healthcare Engineering*, **2019**, 1-12, (2019).

[6] MessanKomi, “*Application of data mining methods in diabetes prediction*”, in Proc. 2nd International Conference on Image, Vision and Computing (ICIVC), 1006-1010, (2017).

[7] M. Renuka Devi and J. Maria Shyla, ”Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus”, *International Journal of Applied Engineering Research (IJAER)*, **11**, 1, 727-730, (2016).

[8] Donna Giri and U.Rajendra Acharya, ”Automated diagnosis of coronary artery disease affected patients using LDA PCA ICA and discrete wavelet transform”, *Knowledge-Based Systems*, **37**, 274-282, (2013).

[9] AyşegülUçar, A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering, *Neural Computing and Applications*, **27**, 1, 131-142, (2016).

[10] GeshwareeHuzooree, *Glucose prediction data analytics for diabetic patients monitoring*, in Proc. 1st International Conference on Next Generation Computing Applications (NextComp), 188-195, (2017).

[11] E. Aboufadel, Robert Castellano and Derek Olson, Quantification of the variability of continuous glucose monitoring data, *Algorithms*, **4**, 1, 16-27, (2011).

[12] R. Bunescu, Blood Glucose Level Prediction Using Physiological Models and Support Vector Regression, *IEEE Machine Learning and Applications*, **1**, 135-140, (2013).

[13] Henry R and Diem P, *Multi-model data fusion to improve an early warning system for hypo-/hyperglycemic events*, in Proc Annu Int Conf IEEE Eng Med Biol Soc , 4843–4846, (2014).

[14] A. Gani and Andrei V, Universal Glucose Models for Predicting Subcutaneous Glucose Concentration in Humans, *IEEE Transactions on Information Technology in Biomedicine*,**14**, 1, 157-165, (2010).

[15] M. P. Reymann, *Blood Glucose Level Prediction based on Support Vector Regression Using Mobile Platforms*, in Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2990-2993, (2016).
[16] E. Daskalaki, Real-Time Adaptive Models for the Personalized Prediction of Glycemic Profile in Type 1 Diabetes Patients, *Diabetes Technology and Therapeutics*, **14**, 2, 168-174, (2012).
[17] K. Zarkogianni, *Neuro-Fuzzy based Glucose Prediction Model for Patients with Type 1 Diabetes Mellitus*, in Proc. IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 252-255, (2014).
[18] M. Eren-Oruklu, Estimation of Future Glucose Concentrations with Subject-Specific Recursive Linear Models, *Diabetes Technology and Therapeutics*, **11**, 4, 243-253, (2009).
[19] Sara Belaroucia and Mohammed Amine Chikh , Medical imbalanced data classification, *Advances in Science, Technology and Engineering Systems*, **2**, 3, 116-124 (2017).
[20] C. Bayraktar, Diagnosing diabetes using neural networks on small mobile devices, *Expert Systems and Applications*, **39**, 1, 54-60, (2012).
[21] K.Ananthajothi, M.Subramaniam, Multi level incremental influence measure based classification of medical data for improved classification, *Cluster Computing* (2018). https://doi.org/10.1007/s10586-018-2498-z
[22] P. Laura Juliet, An Improved Prediction Model For Type 2 Diabetes Mellitus Disease Using Clustering And Classification Algorithms, *International Research Journal of Engineering and Technology (IRJET)*, **6**, 2 , 1179-1186, (2019).
[23] CristinaBianchi, et. al.,Freestyle Libre trend arrows for the management of adults with insulin-treated diabetes: A practical approach, *Elseveir, Journal of Diabetes and its Complications*, **33**, 1, 6-12, (2019).

K. Ananthajothi is an Assistant Professor in the Department of Computer Science and Engineering at Misrimal Navajee Munoth Jain Engineering College, Chennai,(INDIA). He obtained his Master degree (M.E) in Computer Science andEngineering in Anna University in the year 2010 and Pursuing Ph.D in Anna University. His research focuses are Data Mining and Big-data. He published book title of Theory of computation.

M. Subramaniam is a Professor and Head for the Department of Information Technology at S.A. Engineering College affiliated to Anna University, Chennai, (INDIA). He obtained his Bachelor's degree (B.E) in Computer Science and Engineering from University of Madras (1998), Master degree (M.E) in Software Engineering and Ph.D from College of Engineering Guindy (CEG), Anna University Main Campus, Chennai -25 in the year 2003 and 2013 respectively. His research focuses are Computer and Mobile Networks, Cloud, Big-data and Software Engineering. He is an active life member of the Computer Society of India (CSI) and the Indian Society for Technical Education (ISTE). He has six Research scholars perusing Ph.D under his guidance. He published many research papers in reputed journals. He is also reviewer in IEEE- International Journal of Communication Systems.