J Biochem Tech (2019) Special Issue (2): 63-68
ISSN: 0974-2328

A Proposed Method of Feature Weighting to Improve Predictions of Diabetes Mellitus

Hojjat Ahmadinejad, Ahura Ahmadi*, Farzaneh Mosavat, Ali Yousefi

Received: 20 September 2018 / Received in revised form: 10 March 2019, Accepted: 21 March 2019, Published online: 25 April 2019

Abstract

Inability in diabetes mellitus diagnosis in early stages of the disease is one of the main problems for diabetic patients. Hence, in this study we aim to improve the ability of diabetes detection by applying a combination of a classification method as a basic method and k-nearest neighbor and four clustering combination of basic models. The process of weighting features were repeated five times and each step, showed a progress in accuracy of results. Because of too many null data we have used a set of averaging methods, instance removal and a similar method to k-nearest neighbor algorithm in data preparation process.

Key words: Diabetes mellitus, Clustering, K-nearest neighbor.

Introduction

Diabetes is a continuing health challenge for the 21st century. It has a large burden of disease in various populations and almost half of all deaths attributable to high blood sugar occur before the age of 70. WHO demonstrated that diabetes will be the 7th leading cause of death in 2030 (WHO, 2016; Mathers and Loncar, 2006). Despite its high prevalence, till now there is no known eradication method worldwide. The main challenge about diabetes is late diagnosis of this disease. Therefore implementing any method, risk assessment tools and their various algorithms to assist in correctly diagnosis, especially in early stages would be a significant progress to prevent various complications caused by diabetes mellitus (Najafipour et al., 2014).

Several intelligent methods have been demonstrated to overcome this obstacles by now such as: pattern recognition fuzzy algorithms for feature extraction (Chen and Chen, 2002) and Bayesian based methods (Langseth and Nielsen, 2006). It has been established that groups of classifier are able to provide more highly accurate results (Ubeyli, 2009). Dogantekin et al. performed a study using LDA along with artificial neuro FIS (ANFIS) for the detection of diabetes (Dogantekin, Dogantekin and Avci, 2009). As a thumb of rule in data mining; any method has its advantages and disadvantages (Tulyakov et al., 2008). In other words: there is no single method which is the best in all systems. That is why approaches to combine intelligent methods are developing to find a stable and perfect solution (Shankaracharya et al., 2010). This fact is vivid in study of Geloven et al (2002); in which 4 methods are mixed together to construct response model in direct marketing. The purpose of this study is to propose a method of feature weighting to improve predictions of diabetes mellitus

Data Preparation

The data in this study are from dataset related to diabetes named PIDD. The gathered data have 8 features about 500 healthy women and

Hojjat Ahmadinejad
MSc. Engineer of Information Technology, CEO of NouAndish Pars Co., Tehran, Iran.

Ahura Ahmadi*
Assistant Professor, Shahid Beheshti University of Medical Sciences, School of Medical Education, Tehran, Iran.

Farzaneh Mosavat
Assistant professor of oral and maxillofacial radiology Department, School of Dentistry, Tehran University of Medical Science, Tehran, Iran.

Ali Yousefi
MSc. Engineer of Computer Engineering, Department of Computer Engineering, Malayer Branch, Islamic Azad University, Malayer, Iran.

***Email:** Ahura.ahmadi@sbmu.ac.ir

268 type 1 diabetic women older than 21 years which is completely similar in form of indexes and standards of the world health organization.

The features used in this study include:

1. Number of times pregnant
2. Two hour post prandial blood sugar test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. Two-Hour serum insulin (mu U/ml)
6. Body mass index (BMI) (weight in kg/(height in m)^2)
7. Diabetes pedigree function (DPF)
8. Age (years)

After investigating data set so many data are determined as null data (Table 1).

Table 1: Null data in 268 samples

Feature no.	1	2	3	4	5	6	7	8
No. of null data	0	5	35	227	374	11	0	0

According to null data in features 4 and 5, we identified null data in both of them and removed them. 541 remaining data are explained in table 2.

Table 2: Null data in 541 remaining samples

Feature no.	1	2	3	4	5	6	7	8
No. of null data	0	5	2	0	147	2	0	0

Null data in features 3, 2 and 6 are replaced by their average due to their low number, but the fifth feature is still containing too many null data. So, it does not sound logic to be replaced by the average value and another method must be selected. For reaching this goal we used identifying neighbors method in k-nearest neighbor (fig. 1); we divide data set to two separate sets according to null or non-null data of the fifth feature of any sample and then the distance of existing samples in other set based on Euclidean distance for any sample x which has null data was calculated. Finally we identify 15 samples with least distance with x (15-nearest neighbor), replace null data of x by the fifth feature's value.

Input: an attribute-valued dataset D
For all instance x in D do
If 5^{th} feature of x is NaN then
Attach x to dataset A else
Attach x to dataset B
End if
End for
For all instance y in A do
Compute distance between y and all instances in B
Find 15 nearest instance in B
Compute mean of 15 instances and replace to y
End for

Figure 1: The neighbor identifying method in k-nearest neighbor

Basic method configuration

A simple method for classification of patients and healthy groups is using the k-nearest neighbor method. In this method, each class for an instance, the status of the class of "k" for the instance among all of the dataset with the most similarity most be determined. Mixing this method with the clustering methods is for more accuracy. In this study, for each train data, numbers of k-means clustering with large numbers of clustering sets were developed. Afterward regard to the label for the train data, proportion of patient and healthy group for

each cluster was determined. For every instance in train data cluster related to each clustering was recognizing, the mean portion for that cluster for the probability of being patient, for the instance. Now for every instance in the test data k-nearest neighbor method was performed on the train data and the mean of being patient probability among the nearest and most similar "k" to that instance to the probability of the instance being patient.

The logic for adding numeral clustering with mentioned k-means neighbor method decreases the risk of an incorrect diagnosis in a train data. Furthermore, instead of using a binary label, the usage of percentage for clusters, lead to a ranking, which improves the decision making process.

Process of the suggested weighting method

1. General Points of weighting Model:

Here we generally discuss about diagnosis between to patient and healthy groups to score one point to any data sample by implementing basic method. This score shows the probability that data sample belongs to any group from the method's point of view. Therefore, we try to combine the obtained scores by prepare weighting for input data in a way to improve diagnosis.

Diabetic group are identified by label "1" and healthy group are labeled by "zero", in this data set. Hence people based on the probability of being diabetic in the applied algorithm's point of view can be identified, by ranking the obtained scores resulted from implementing algorithms.

In the mentioned method, just one the features of the weighing inputs are being evaluated in any step. According the best value for the current features are obtained by considering other features as constant. Then similar operations are executed for other features until all features are being adjusted. This process can be done for several times and improve results on training data.

2. Technical Statement of The weighting Model:

At first with the min-max method all the input features were normalized. So for any n input features and m instance we have:

$$\begin{bmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,n} \\ S_{2,1} & S_{2,2} & \cdots & S_{2,n} \\ \vdots & \vdots & & \vdots \\ S_{m,1} & S_{m,2} & \cdots & S_{m,n} \end{bmatrix}$$

Where $S_{a,b}$ is the feature b^{th} of a^{th} instance .In this weighting method, the basic algorithm scores all of the instances. Therefore the basic algorithm regard to m number of instances, the vector includes the values of T_1, T_2, T_3 … T_m which each of the instances has a value between zeroes to 100.

Based on this, given W_1, W_2… W_n as the weight of features, we have:

$$S_1 = S_{1,1}W_1 + S_{1,2}W_2 + S_{1,3}W_3 + \ldots + S_{1,n}W_n$$

$$S_2 = S_{2,1}W_1 + S_{2,2}W_2 + S_{2,3}W_3 + \ldots + S_{2,n}W_n$$

$$S_3 = S_{3,1}W_1 + S_{3,2}W_2 + S_{3,3}W_3 + \ldots + S_{3,n}W_n$$

$$\vdots$$

$$S_m = S_{m,1}W_1 + S_{m,2}W_2 + S_{m,3}W_3 + \ldots + S_{m,n}W_n$$

Where S_1, S_2, … , S_m are the scores of samples by the basic algorithm, based on weights of the basic algorithms. The aim is to score S_1, S_2, … , S_m in such a way that T_1, T_2, T_3, …, T_m equivalent to them has more accuracy.

Now in order to weight features, we assume the weight of all features as constant except one (k^{th} feature) and obtain the best value for

W_k. Repeating the process the optimal weight of all n existing features are calculated.

In Equation $S = S_1W_1 + S_2W_2 + S_3W_3 + \ldots + S_kW_k + \ldots + S_nW_n$ for W_k as variable and other Ws as constant, the expression $S_1W_1 + S_2W_2 + S_3W_3 + \ldots + S_{k-1}W_{k-1} + S_{k+1}W_{k+1} + \ldots + S_nW_n$ would be constant which is called B and also let $A=S_k$.

Therefore by combining the score of the samples of n features whose weights are constant except K^{th} feature, the new scored values are obtained as follows:

$S_1 = W_kA_1 + B_1$

$S_2 = W_kA_2 + B_2$

$S_3 = W_kA_3 + B_3$

⋮

$S_m = W_kA_m + B_m$

In which $S_1, S_2, S_3, \ldots, S_m$ are one-degree equations with W_k as variable, since A and B are constant.

Any 2 equations will have one solution (crossing point) if they do not have equal slope (A values) which is obtained as follows:

$y = a_1x + b_1$, $y = a_2x + b_2$

$t = (b_1 - b_2) / (a_2 - a_1)$

Figure four, shows 4 equations with hypothetical values which have 6 crossing points since their slopes are not identical.

By assuming $W_k=0$ initially, ordering instances $S_1, S_2, S_3, \ldots, S_m$ which are equal to $B_1, B_2, B_3, \ldots, B_m$ respectively ordered. The crossing point of any 2 equations is a threshold (t) which switches the rank of two relative instances. If both instances are patient or healthy group, this switch does not result in any change, but if one of them is patient or healthy group, then one score is changing in area under diagram.

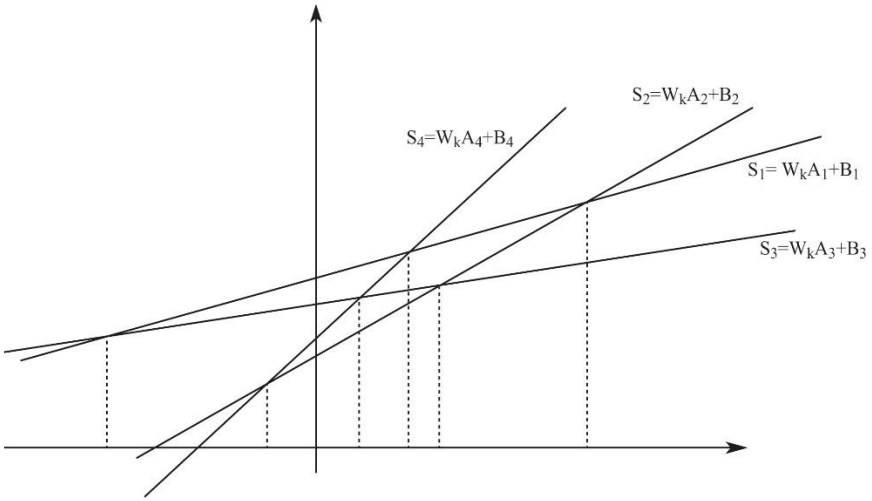

Figure 2: Four equations on the diagram

Therefore, by reviewing the label of the train data and separating patient and healthy group instances we find the crossing points of all patient instances with healthy group ones and obtain their threshold. If the crossing point belongs to $(-\infty,0)$ on x axis, it is ignored, since we do not consider negative weights, but if it is in $(0,+\infty)$ on x axis, then it is considered as a threshold (t).

By creating *m* threshold we will have *m+1* intervals $[0,t_1],[t_1,t_2],\ldots,[t_m,+\infty]$ and we obtain new scoring (area under diagram) by switching 2 related instances in ranking in any threshold, which has only one different score (area under diagram) between the last and next threshold. Now ordering threshold values, we obtain the area under gain diagram for threshold and find the highest area.

Figure 3 shows this for 4 hypothetical equations in figure 2. The interval $[t_2,t_3]$ is the best in this example.

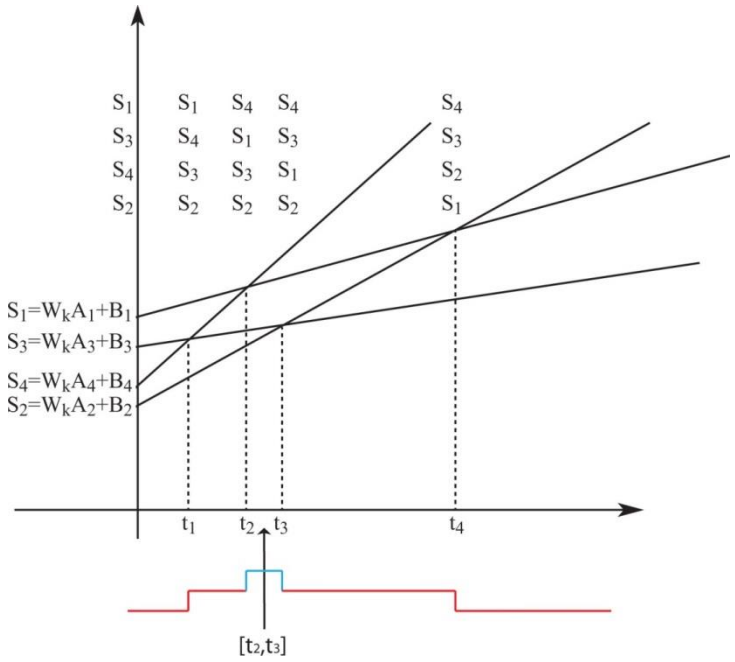

Figure 3: Best interval for weighting with 4 given samples

W_k is assumed to be equal to 0, if $[0,t_1]$is the best interval, assumed to be t_m+T (T is a sufficiently large number which is assumed to be 10 in this paper) if $[t_m,+\infty]$ is the best one and is equal to the mean of the interval otherwise. The Above process is repeated for all *n* existing values of W in *n* features in order to have a combination of existing features which are weighted. This process is repeated to find better weights.

It should be noted that in next steps in which sample weights are changing, W_k is not remaining as the best weight. Therefore, the weighting algorithm must be repeated multiple times with last adjusted weights as initial weights. In any iteration, classifying ratio decreases or remains unchanged, since we select the best weight for one sample in each step. We can repeat this until a great improvement.

Analyzing the findings

The applied data set after data preparation step contains 541 samples and each of them has 8 features labeled by 0-1 values for healthy/patient group. The samples were selected randomly with ration of 75% for train and 25% for test in order to implement the basic algorithm (405 training instances and 136 test instances).Afterwards all input data features were normalized according to min-max method.

With determining the frequency of cluster with values 80, 100, 120, and 150, four clustering with k-means method on train data were performed and regard to the means سهم each of the four clusters according to the instances, percentage of being patient according to instances in train data were calculated. Meanwhile the weights of all the input features were equal one.

For weighting input features according to mentioned method, in every stage seven of the features were constant while the weight of the one remaining feature was adjusted. Weighting all features were repeated five times, coefficients can be seen in table 3.

Table 3: The obtained data for five iterations of basic methods

Measured weights for input features with 5 iterations								
	P 1	P 2	P 3	P 4	P 5	P 6	P 7	P 8
1 iteration	1/202	2/4536	0/004	0/3973	0/0014	2/6241	1/2987	1/8199
2 iteration	0/4708	2/9635	0/0007	0/4635	1/6189	1/5495	1/3189	0/3231
3 iteration	0/9942	1/9576	0	0/3318	0/0009	2	1/6963	2/6333

4 iteration	0/4929	4/1435	0	0/0027	0/5629	1/7726	1/186	1/0574
5 iteration	0/1734	4/2938	0/0024	0/1561	0/6633	2/7665	1/4163	1/211

For assessing the effect of performed weighting, all if the five groups' weights were performed on the test data then with the threshold equal to 50% were compared to each other. The results can be seen in table 4.

Table 4: Performance indicators.

No weight: TP= 32, FP= 20, TN= 71, FN= 13 1 iteration: TP= 37, FP= 16, TN= 75, FN= 8 2 iteration: TP= 36, FP= 14, TN= 77, FN= 9 3 iteration: TP= 40, FP= 14, TN= 77, FN= 5 4 iteration: TP= 40, FP= 11, TN= 80, FN= 5 5 iteration: TP= 33, FP= 13, TN= 78, FN= 12							
Performance indicators	Explanation	No weight	1 iteration	2 iteration	3 iteration	4 iteration	5 iteration
False positive rate	$\frac{FP}{Total\ number\ of\ instances}$	14.71%	11.76%	10.29%	10.29%	8.09%	9.56%
False negative rate	$\frac{FN}{Total\ number\ of\ instances}$	9.56%	5.88%	6.62%	3.68%	3.68%	8.82%
True positive rate	$\frac{TP}{TP+FN}$	71.11%	82.22%	80%	88.89%	88.89%	73.33%
Agreement rate (accuracy)	$\frac{TP+TN}{Total\ number\ of\ instances}$	75.74%	82.35%	83.09%	86.03%	88.24%	81.62%

Regard to table 4 it can be obtained that accuracy with iteration weighting have been improved and in the fourth iteration the best result was obtained whilst in fifth iteration a deteriorated result was achieved, this can be due to over fitting in train data.

Conclusion

Overall in this study, features were weighted based on the impact they have on diagnosis on the diabetes; in such a manner that it can lead to better diagnosis of the disease. According to this method, in every stage with concentration on a feature and determining its weight factor, the efforts are to improve ranking of instances based on probability of being unhealthy. In this research, the basic method used for classification of instances includes; mixing a k-nearest neighbor and four clustering based on k-means method. The outcomes showed that the suggested method after four step of feature weighting reached to the best results, in a manner that the accuracy value for weighted test data was 88.24%.

References

Chen J, Chen C. (2002). Fuzzy Kernel perceptron. Journal of Neural Networks, IEEE.13(6):1364 – 73.

Dogantekin E, Dogantekin A, Avci D. (2009). An intelligent diagnosis system for diabetes on Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System: LDA-ANFIS Digit Signal Process. 20:1248–55.

Geloven S. (2002). Combining Target Selection Algorithms in Direct Marketing. Delft University of Technology Dutch.

Langseth H, Nielsen T. (2006). Classification using Hierarchical Na¨ıve Bayes models. Journal of Machine Learning. 63(2):135-59.

Mathers C, Loncar D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med. 3(11):e442.

Najafipour F, Ahmadinejad H, Norouzi A, Ahmadi A. (2014). A Combinatorial Method For Diabetes Diagnosis. SAJMR. 3(1):e pub.

Shankaracharya, Odedra D, Samanta S, Vidyarthi AS. (2010). Computational Intelligence in Early Diabetes Diagnosis: A Review. Rev Diabet Stud. 7(4):252–62.

Tulyakov S, Jaeger S, Govindaraju V, Doermann D. (2008). Review of Classifier combination Methods. Studies in Computational Inteligence (SCI). 90:361-86.

Ubeyli E. (2009). Modified mixture of experts for diabetes diagnosis J Med Syst. 33:299–305.

World Health Organization. (2016). Global report on diabetes: World Health Organization.