An intelligent system for diabetes prediction

Conference Paper · June 2015

DOI: 10.1109/MECO.2015.7181948

3 authors:

Zhilbert Tafa
University for Business and Technology
24 PUBLICATIONS **380** CITATIONS
SEE PROFILE

Nerxhivane Pervetica
Universum College
3 PUBLICATIONS **45** CITATIONS
SEE PROFILE

Bertan Karahoda
University for Business and Technology
7 PUBLICATIONS **49** CITATIONS
SEE PROFILE

An Intelligent System for Diabetes Prediction

Zhilbert Tafa
Dep. of Computer Science
University for Business and Technology
Prishtina, Kosovo
tafaul@t-com.me

Nerxhivane Pervetica
Bertran Karahoda
Dep. of Computer Science
University for Business and Technology
Prishtina, Kosovo
np32972@ubt-uni.net
b.karahoda@ubt-uni.net

***Abstract*—With the emerging increase of diabetes, that recently affects around 346 million people, of which more than one-third go undetected in early stage, a strong need for supporting the medical decision-making process is generated. A number of researches have focused either in using one of the algorithms or in the comparisons of the performances of algorithms on a given, usually predefined and static datasets that are accessible through the Internet. This paper focuses on the joint implementation of the support vector machine (SVM) and Naïve Bayes statistical modeling, in the dataset acquired from the medical examinations of 402 patients, in order to improve the computer-supported diagnosis reliability. The dataset contains some attributes that have not been previously used in computer-based evaluations. The results show that the joint implementation of two algorithms improves significantly the overall reliability of the system outcome, which is crucial in the computer-supported diabetes diagnostic process.**

Keywords- algorithms; diabetes; joint implementation, machine learning; Naïve Bayes; support vector machine.

I. INTRODUCTION

Diabetes is a chronic disease caused by the increase in blood sugar, mainly either due to the less production or no production of insulin in body (type 1diabetes), or due to the fact that cells do not respond to the produced insulin (type 2 diabetes). In recent years, the number of diabetic patients has increased drastically, as noted in [1], mainly due to the aging population and irregular western food habits. According to the World Health Organization, diabetes affects around 346 million people in the world, with the prevalence of diabetes type 2. Moreover, diabetes is the major cause for heart stroke, kidney failure, lower-limb amputations and blindness. As presented in [2], the absence of symptoms, or the absence of recognition of the indicators in the patient's data, may lead to the pre-diabetes or diabetes condition that goes undetected even in more than one-third of people that are later diagnosed with diabetes.

During the clinical examinations of various forms, lots of data are acquired from patients. The development of the computer-based methods that would enable the high probability recognition of pre-diabetic or diabetic condition can be an efficient support to the decision making in healthcare.

Machine learning is the area of artificial intelligence that uses the statistical analyses, and is recognized to be a promising area that, based on the given dataset of diabetics, can help in patient classification or probability prediction regarding the patient's pre-diabetic or diabetic condition. The main strength of these methods is contained in the ability of the algorithms to learn from data and to use that knowledge for later predictions and decisions. There are a number of machine learning and statistical modeling approaches that so far have been involved in various aspects of solving the problem. According to [3], although other classifiers perform well, the SVM outperforms other classifiers with respect to accuracy, sensitivity, specificity, and precision.

This paper presents a joint Matlab implementation of the SVM and Naïve Bayes methods in a new dataset acquired from the patients examined for diabetes in Kosovo. The developed diagnostic tool enables the intelligent computer-based prediction on diabetes, based on the previously acquired values. The statistical analysis shows the high accuracy of data classification. Also, the proposed joint implementation of two algorithms aims to improve the reliability of the decision by using the power of both algorithms in minimizing their individual weakness.

The rest of the paper is structured as follows. Related work on the topic is presented in Section II. In section III, the experimental setup and the implementation are presented. Results are shown and discussed in Section IV while the Section V concludes the work.

II. RELATED WORK

A number of research efforts have been directed in involving the machine learning algorithms to the design of intelligent healthcare applications, especially in disease detection. Most of them have focused on heart disease, cancer detection, and diabetes. An analytical study of several algorithms, focused on classification of diabetes mellitus data, is presented in [1]. The algorithms have mostly focused on detection of pre-diabetes, which was recognized in [4] as a relatively strong indication for the future development of diabetes. A recent study in this direction is given in [5], where two machine learning techniques, namely SVM and ANN

(Artificial Neural Networks), are used to predict pre-diabetes in Korean population. A similar approach is used in studying the correlation for hematological parameters and glucose level for identification of diabetes [6].

In general, the research focus is to conduct some of the supervised learning algorithms on the given dataset and extract the knowledge about the prediction of diabetes based on given values of the appropriate attributes. In [7], the SVM implementation gives the prediction accuracy of 94%. Another implementation of the SVM in detecting the diabetes is given in [8]. Here, the SVM classifier, however, performs only 78 % of accuracy. A method for prediction of diabetes by using Bayesian network is given in [9] while the authors in [10] separately use Naïve Bayes and k-nearest neighbor algorithm.

Most of the mentioned researches, rely on Pima Indian database of diabetic, and therefore have the same attributes and similar conclusions. Furthermore, they treat one or two algorithms independently to compare the efficiency of the algorithms between each other. Some studies, however, recommend the hybrid use of a distance-based algorithm and a statistical based method [11] or the combination of classification and clustering [1].

In contrast to the most of the mentioned materials and methodologies, this paper treats original dataset, with different attributes, extracted from the medical examinations in Kosovo. Beside the algorithm performance evaluations, the aim of the paper is not only to analyze the dataset and to provide a supporting tool for diabetes detection, but also to improve the derived decision reliability by jointly using two machine learning algorithms. This creates a more reliable zone (with the answers yes/no on diabetes) and a “grey zone” that would direct the decision making process to the further clinical examinations.

III. MATERIALS AND IMPLEMENTATION METHODS

A. Dataset description

Dataset consists of 402 instances taken from three different locations in Kosovo. During the data acquisition process, the appropriate importance is given to the patient’s data privacy and anonymity. The attributes of the database are: BMI (body mass index), glucose level before meal and after meal, the systolic and diastolic blood pressure, the hereditarily factor, the regular diet, and daily physical activities. The last two attributes are evaluated as follows. Regarding the issue of regular diet, while relying on inputs from the medical clinicians, patients were asked if they took their meals in approximately same equidistant daily intervals at least three times a day and also if their meals were not voluminous. With these answers being positive, we consider that a patient is having the regular diet. On the other hand, according to the U.S. Center for Disease Control and Prevention (CDCP), the adult person is considered to be physically active if he/she conducts the 150-200 min of physical activities a week. With the answers on the family history questions, and in accordance to the above given thresholds, the answers of the examinee regarding the last three questions are mapped into two values: 1 and 0.

The ranges of the values of all attributes are given in Table I.

TABLE I. THE RANGES OF THE ATTRIBUTES

Attribute	Value range	
	From	***To***
BMI	15	40
Pre meal glucose	3.5	19
Post meal glucose	4.9	22.8
Diastolic blood pressure	55	110
Systolic blood pressure	90	200
Family history of diabetes	No (0)	Yes (1)
Regular diet	No (0)	Yes (1)
Physical activities	No (0)	Yes (1)

After acquiring the given initial data from patients, and after the extensive laboratory examinations and continuous monitoring, 80 of patients were diagnosed with type 2 Diabetes. The presence of diabetes in an instance was labeled with Boolean 1. The rest of patients were not diagnosed with diabetes, which results in number of 322 Boolean zeros in the dataset.

B. The SVM implementation

The SVM algorithm represents the instances as points in space, mapped so that separate classes are divided by a clear gap. The aim is to find the maximum-margin hyper plane – the one that gives the greatest separation between the classes. The instances that are closest to the maximum-margin hyper plane are called support vectors. Support vectors are chosen based on the portion of the dataset that represents the training set. Support vectors of two classes enable the creation of two parallel hyper planes. The larger the margin between these two hyper planes, the better the generalization error of the classifier.

Training data points can be represented in form:

{(X1, Y1), (X2, Y2)…., (Xn, Yn)}

where Xi is a k-dimensional vector and Yi is +1 or -1 denoting the class to which a given point belongs to. The training data is then divided by a hyper plane of general form:

$$W * X + B = 0 \tag{1}$$

Where W is k-dimensional vector, perpendicular to the hyper plain and B is scalar. Two parallel hyper planes that belong to two different classes can be described by equations:

$$W * X + B = 1 \tag{2}$$

$$W * X + B = -1 \tag{3}$$

The distance between the hyper planes is 2/|W| so the aim is to minimize |W|. The semantics behind the multidimensional formulations (1)-(3) is given in [12] and further process of minimization of factor 2/|W| is explained in [13].

C. The Naïve Bayes implementation

The Naïve Bayes classification is based on the probabilistic Bayes theory. As noted in [11], the Naïve Bayes statistical algorithm is a frequently used method in prediction problem. The implementation of the statistical modeling is based on linear function. Theoretically, it usually means the appliance of unrealistic assumption that the attributes are equally important and independent. The real life dataset consists of attributes that are certainly not equally important or independent, but, as noted in [12] and also shown in this paper, it leads to a simple scheme that, again, works surprisingly well in practice.

While keeping in mind the above mentioned independency assumption, the calculation of the probability that a given record belongs to class Y=C, can be calculated as the product of probability that each of values of the *i* record's attributes belong to class C, i.e.,

$$P(X|Y=C) = \prod_{i=1}^{q} P(X_i|Y=C) \tag{4}$$

The probability that a given value of the attribute belongs to class y_j, when the dataset contains numerical inputs, can often generally be calculated by using the Gaussian distribution function, i.e.,

$$P\left(X_i = x_i \middle| Y = y_j\right) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{\left(x_i-\mu_{ij}\right)^2}{2\sigma_{ij}^2}} \tag{5}$$

Finally, the probability that a given record will be classified in class C, can be formulated with:

$$P(C|x_1, \dots x_n) = \frac{P(C)P(x_1,\dots,x_n|C)}{P(x_1,\dots x_n)} \tag{6}$$

D. The proposed architecture

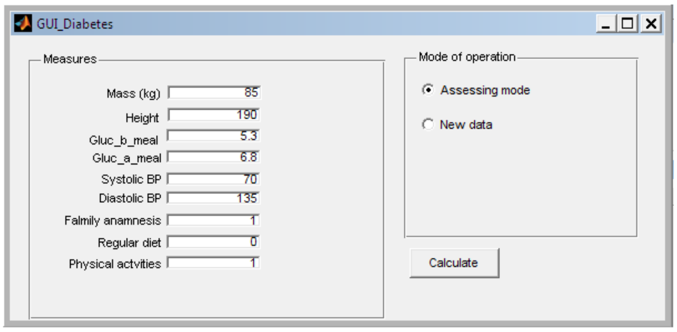
The system consists of the following elements. Two machines, namely SVM and Naïve Bayes classifier, build their classifiers based on the training sets. Afterward, they are ready to perform the classification of a given record. The output can belong to one of two classes: class 0 (no diabetes) and class 1 (diabetes). If the outcomes from both classifiers are equal, the record is classified as belonging to class 0 or class 1. If the output is different, then the record is considered as still unclassified (grey zone).

The SVM algorithm is implemented by using bioinformatics tool in Matlab while Naïve Bayes implementation is constructed manually, also in Matlab. Matlab was chosen due to its flexibility and ability to work with various file formats. By using the Matlab function, the data are divided into the training set and the testing set. The classifier is extracted from the training set while the instances from the testing set are tested on the derived classifier. The SVM algorithm uses the polynomial kernel. In order to mitigate any bias caused by samples chosen for holdout, the repeated holdout method of error estimation is used. In this direction, the process of classifier performance evaluation is repeated 100 times with the classifier performance evaluated in each of the iteration.

IV. RESULTS AND DISCUSSIONS

Prior to building the application, analyzing data, and evaluating the classification performance, sample size is conducted to analysis in accordance to the methodology described in [14]. With the given number of population, the given response distribution of diabetes in Kosovo, and the given confidence level of 98%, the margin of error that occurs due to the limited number of examinee is expected to be in the range of 1,62 %.

The developed Matlab-based application works in two modes: the assessment mode and the data acquisition mode. The GUI is given in Fig. 1.



GUI_Diabetes

Measures

Mass (kg) 85

Height 190

Gluc_b_meal 5.3

Gluc_a_meal 6.8

Systolic BP 70

Diastolic BP 135

Falmily anamnesis 1

Regular diet 0

Physical actvities 1

Mode of operation

● Assessing mode

○ New data

Fig. 1: Diabetes diagnostic application GUI

The data acquisition mode enables for the addition of new records, in order to increase the sample size and the classification accuracy. The assessment mode gives the prediction of pre-diabetic or diabetic condition on each of the newly added record.

The prediction on new record is based on the execution of both SVM and Naïve Bayes classification on previous instances with different randomly chosen training sets. If both classifiers classify the record as being positive or negative, then we consider the given output of the classifier for a specific record as having high reliability. Otherwise, if the outputs are different, the patient needs to be further monitored on diabetes.

For the simplicity reasons, by incorporating the stratification, the data are quasi-randomly split into two equal sets - 201 instances for training and 201 instances for evaluating the performances. In general, these two sets are only approximately equal, since the numbers of instances for training and testing are selected automatically by *crossvalind* Matlab function. In order to find the average value of the classifiers' accuracy, the process of random selection of the training set and test set along with the classifier performance evaluation on each random selection is repeated 100 times.

The results show the mean value of the SVM classifier performance - accuracy of 95, 52 % while for the Naïve Bayes classifier the classifier accuracy is 94, 52%. Both values vary in +/- 1% of classification performance margin during various iterations. This also shows for the high stability of classifier.

The average number of correctly and incorrectly classified records is calculated for both classifiers and the results are given in Table II along with the structure of the sample distribution for training and testing the system.

TABLE II. CLASSIFICATION AND THE AVERAGE ACCURACY

	No. of records	Train set / test set	No. of correctly classified	No. of incorrectly classified records	Classifier performance (mean value)
SVM	402	201/201	192	9	95.52%
Naïve Bayes	402	201/201	190	11	94.53%

As can be noted, and as expected, in terms of the overall average classification accuracy, SVM over performs the Naïve Bayes classification, but the difference in classifier performance is surprisingly small.

The performances are also estimated in terms of other multi-class classification measures such as precision and recall. For a specific class, these two measures are calculated as follows:

$$Precision_X = \frac{TP_X}{TP_X + FP_X} \quad (7)$$

$$Recall_X = \frac{TP_X}{TP_X + FN_X} \quad (8)$$

The notations TP, FP and FN refer to the number of data classified as true positive, false positive and false negative, i.e., the data correctly classified in class X, the data that belong to another class and are incorrectly classified in class X, and the data that should have been classified in class X whereas are classified in another class..

The results regarding the precision and recall of classes YES and NO are given in Table III.

TABLE III. PRECISION AND RECALL

	Precision (Class NO)	Recall (ClassNO)	Precision (Class YES)	Recall (Class YES)
SVM	0.97	0.975	0.892	0.868
Naïve Bayes	0.981	0.951	0.814	0.921

The joint implementation aims to improve the reliability of the decision in case when both algorithms give the same outcome, when we consider the outcome to be valid. Otherwise, when the outcomes are different, we consider the decision to be unreliable and hence invalid.

After the repeated joint implementation of both algorithms, the average rate of the valid outcomes turns out to be 94, 77 % with the expected lower bound of approximately 90, 3 %. This means that the major number (94, 77%) of new patients that need to be classified regarding the likeness of diabetic condition, will be classified the same way by two independent classifiers. The difference of 5.23 % in the results of classifiers exists due to their individual weaknesses and can also be related to the dataset. Among the valid outcomes, the derived accuracy of the decision is now improved up to 97,6 %, which significantly over performs both classifiers individually.

V. CONCLUSIONS AND FUTURE WORK

The research efforts presented in this paper are focused in developing and the evaluation of a computer-based support tool for the diabetes detection.

The presented approach is based on the joint implementation of two algorithms in Matlab that have been executed on the newly acquired dataset with the different attributes as compared to the previous work in this field. The algorithms are executed and evaluated independently but the decision making is based on the joint outcomes from both algorithms. The aim of this approach is to make the decision more reliable.

As shown in the paper, both SVM and naïve Bayes algorithm have individually shown high overall classifier performances of 95, 52% and 94, 52%, respectively. The joint implementation on the same, newly added record leads to one of the three answers: a) the patient is diagnosed with diabetes (or pre-diabetic condition), b) the patient is not diagnosed as having the mentioned condition, and c) the patient is further directed to the additional clinical examinations. If two algorithms show different results, the answer is classified as condition c). Otherwise, the accuracy of the answers a) or b), as shown in the paper, is improved up to the value of 97,6%. The presented methodology minimizes the false negative answers, which is a crucial issue in medical diagnoses.

Finally, the construction approach, the architecture, and the evaluation of a diabetes classification tool presented in this paper, should provide an important guideline to further construction of the similar applications on improving and helping the decision making process in disease detection. The development of a user-friendly and widely accessible application would enable the personal self-screening on diabetic or pre-diabetic condition which is crucial to the disease treatment performance.

The future work will focus on further quantitative evaluations of the developed tool regarding the extensive clinical examinations and results. Also, other methods should be involved in finding the best fit in the sense of accuracy, processing time, etc. The influence of cultural-related biases (such as those related to the nutrition structure and habits) should also be minimized in order for the results to get more generalized.

ACKNOWLEDGMENT

This paper is based upon work in MESI – the project for developing and implementation of the embedded medical applications supported by Montenegrin Ministry of Science. We gratefully acknowledge additional support from University for Business and Technology.

REFERENCES

[1] S. Peter, “An Analytical Study on Early Diagnosis and Classification of Diabetes Mellitus,” *Bonfring International Journal on Data Mining*, vol. 4, no.2, pp. 7-11, 2014.

[2] J. C. Yong, C. K, Hyeon, M. K. Hee, W. P. Seok, J. Jongoh, and J. K. Dae, “Prevalence and Management of Diabetes in Korean Adults: Korea National Health and Nutrition Examination Surveys,” *Dibetes Care*, vol. 32, no. 11, pp. 2016-2020, 2009.

[3] S. Aruna, S. P. Rajagoplan, and L. V. Nandakishore, “An Empirical Comparision of Supervised Learning Algorithms in Disease Detection,” *International Journal of Information Technology Convergence and Services*, vol. 1, no. 4, pp. 81-92, 2011.

[4] American Diabetes Association, “Diagnosis and Classification of Diabetes Mellitus,” *Diabetes Care*, vol. 35, pp. s64-s74, 2012.

[5] S. B. Choi, W. J. Kim, T. K. Yoo, J. S. Park et al.,”Screenng for Prediabetes Using Machine Learning Models,” *Computational and Mathematical Methods in Medicine*, pp. 1-8, 2014.

[6] A. Worachartcheewan, C. nanteasenamat, P. Prasertsrithong, J. Amaranan, T. Monnor et al., “Machine Learning Approaches for Discerning Intercorrelation of Hematological Parameters and Glucose Level for Identification of Diabetes Mellitus,” *EXLI Journal*, pp. 885-893, 2013.

[7] R. Aishwarya, P. Gayathri, and N. Jasinkar, “A Method for Classification Using Macnine Learning Technique for Diabetes,” *International Journal of Engineering and Technology*, vol. 5, no. 3, pp. 2903-2908, 2013

[8] V. A. Kumari and R. Chitra, “Classification of Diabetes Using Support Vector Machine,” *International Journal of Engineering Research and Applications*,” vol. 3, iss. 2, pp. 1797-1801, 2013.

[9] M. Kumari, R. Vohra, A. Arora,”Prediction of Diabetes Using Bayesian Network,” *International Journal of Computer Science and Information Technologies*, vol. 5 (4), pp. 5174-5178, 2014.

[10] A. N. R. Nurhayati, “Implementation of Naïve Bayes and K-Nearest Neighbor Algorithm for Diagnosis of Diabetes Mellitus,” *Applied Computational Science*, pp. 117-120, 2013

[11] E. Z. Ferdousy, M. Islam, and M. A. Martin, “Combination of Naïve Bayes Classifier and K-Nearest Neighbor (cNK) in the Classification Based Predictive Models,” *Computer And Information Science*, vol. 6, no. 3, pp. 48-56, 2013.

[12] I. H. Witten, E. Frank, M. A. Hall, “Algorithms: The Basic Methods” in *Data Mining: Practical Machine Learning Tools and Techniques*, third ed., Burlington MA, 2011, ch.4, sec. 4.1, pp. 86-87.

[13] D. K. Srivastava & L. Bhambhu, “Data Classification Using Support Vector Machine,” Journal of Theoretical and Applied Information Technology, pp. 1-7, 2009.

[14] J. E. Bartlett, J. W. Kotrlik, and C. C. Higgins, “Organizational research: Determining appropriate sample size in survey research,” *Information Technology, Learning and Performance. Journal.*, vol. 19, no. 1, pp. 43–50,2001.