SVM based Diabetic Classification and Hospital Recommendation

Shital Tambade
Pimpri Chinchwad College of Engineering, Pune, India

Madan Somvanshi
Pimpri Chinchwad College of Engineering, Pune, India

Pranjali Chavan
Pimpri Chinchwad College of Engineering, Pune, India

Swati Shinde, PhD
Pimpri Chinchwad College of Engineering, Pune, India

ABSTRACT
Today Diabetes has become a severe disease that is growing rapidly worldwide. A lot of research and work has been done on the same and it shows that there is a need of some automated system which would help the diabetic patients to get hospital recommendations and all.The proposed system uses the SVM classifier to classify the person into diabetic positive or negative class. The diabetic positive patients are then clustered into different clusters as per the severity of the disease. The system also recommends all the nearby hospitals to the patients and the generation of QR code reduces the patients headache of carrying the papers/reports, and thus helps the doctors to better understand the patient's diabetic case history.

General Terms
Diabetic Classification,SVM,Pima Indian Diabetic Dataset.

Keywords
Classification,svm,diabetes,pima indian diabetic dataset..

1. INTRODUCTION
Today Diabetes has become a severe disease that is growing rapidly worldwide. A lot of research and work has been done on the same and it shows that there is a need of some automated system which would help the diabetic patients to get hospital recommendations and all.The proposed system uses the SVM classifier to classify the person into diabetic positive or negative class. The diabetic positive patients are then clustered into different clusters as per the severity of the disease. The system also recommends all the nearby hospitals to the patients and the generation of QR code reduces the patients headache of carrying the papers/reports, and thus helps the doctors to better understand the patient's diabetic case history.[3]

There are two general reasons for diabetes:

(1) The pancreas does not make enough insulin or the body does not produce enough insulin. Only 5-10 % of people with diabetes have this form of the disease (Type-1).

(2) Cells do not respond to the insulin that is produced (Type-2).

Insulin is the principle hormone that regulates uptake of glucose from the blood into most cells (muscle and fat cells). If the amount of insulin available is insufficient, then glucose will not have its usual effect so that glucose will not be absorbed by the body cells that require it. Diabetes mellitus being one of the major contributors to the mortality rate. Detection and diagnosis of diabetes at an early stage is the need of the day. Diabetes disease diagnosis and interpretation of the diabetes data is an important classification problem . A classifier is required and to be designed that is cost efficient, convenient and accurate. Artificial intelligence and Soft Computing Techniques provide a great deal of human ideologies and are involved in human related fields of application. These systems find a place in the medical diagnosis [4-5].

2. MOTIVATION
Diabetes mellitus is one of the most serious health challenges in both developing and developed countries. It has become leading cause of death. Detection of diabetes with optimal cost and better performance is the need of the age. Diabetes disease diagnosis via proper interpretation of the Diabetes data is an important classification problem.There is need of some small storage function to keep data very safe instead of maintaining various documents.QRCode plays important role in reducing the paperwork.Portable QRCode keep important data very safe for longer time.Hospital recommendation gives special service for particular disease to respective patients.[

3. SUPPORT VECTOR MACCHINE

3.1 SVM Model Generation
SVM is a set of related supervised learning method used in medical diagnosis for classification. SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM is called Maximum Margin Classifiers. SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The kernel trick allows constructing the classifier without explicitly knowing the feature space. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible For example, given a set of points belonging to either one of the two classes, an SVM finds a hyperplane having the largest possible fraction of points of the same class on the same plane. This separating hyperplane is called the optimal separating hyperplane (OSH) that maximizes the distance between the two parallel hyper planes and can minimize the risk of misclassifying examples of the test dataset. Given labeled training data as data points of the form.[3]

$$M = \{(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_{n,} y_n)\} \qquad [2]$$

where y_n =1/-1, a constant that denotes the class to which that point x_n belongs. Where,

n=number of data sample.

Each x_n is a p-dimensional real vector. The SVM classifier first maps the input vectors into a decision value, and then performs the classification using an appropriate threshold

value. To view the training data, we divide the hyperplane, which can be described as:

Mapping: $w^T.x + b = 0$

where w is a p-dimensional weight vector and b is a scalar. The vector w points perpendicular to the separating hyperplane. The offset parameter b allows to increase the margin. When the training data are linearly separable, we select these hyperplanes so that there are no points between them and then try on maximizing the distance between the hyperplane. We have found out the distance between the hyperplane as 2/ |w|. To minimize |W| ,We need to ensure for all i either [9]

$$w.x_i - b \geq 1 \text{ or } w.x_i - b \leq -1$$

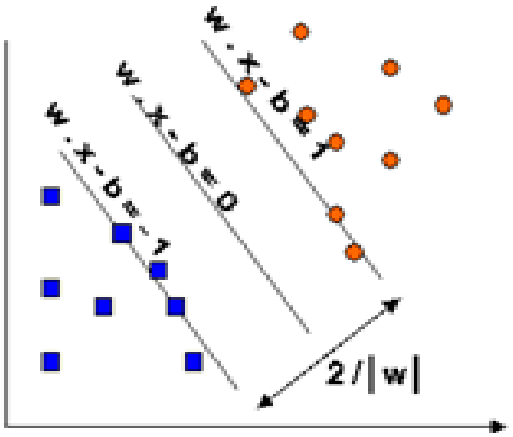

Fig. 1. Maximum margin hyperplanes for SVM trained with samples from two classes

3.2 Radial Basis Kernel Function

The Radial Basis Function (RBF) kernel of SVM is used as the Classifier, as RBF kernel function can analyse higher-dimensional data . The output of the kernel is dependent on the Euclidean distance of from (one of these will be the support vector and the other will be the testing data point). The support vector will be the center of the RBF and will determine the area of influence this support vector ²has over the data space. RBF Kernel function can be defined as [8]

$$k(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

$$\gamma > 0 \qquad [2]$$

where k is a kernel parameter and is the training vector. A larger value of will give a smoother decision surface and more regular decision boundary. This is because an RBF with large will allow a support vector to have a strong influence over a larger area. The best parameter set is applied to the training dataset and the classifier is obtained. The designed classifier is used to classify the testing dataset.

4. K-MEANS CLUSTERING

The Algorithm K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an objective function, in this case a squared error function.[6-7]

The algorithm is composed of the following steps:

Step 1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

Step 2. Assign each object to the group that has the closest centroid.

Step 3. When all objects have been assigned, recalculate the positions of the K centroids.

Step 4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.[2]

5. EXISTING SYSTEM

Practo is a complete health app to book doctor appointments at clinics and hospitals, order medicines set medicine reminders, consult doctors online, manage digital health records & read health tips. It is easiest way to book appointments with doctors, clinics & hospitals. Pick a city, choose a doctor, select a time and done.It provides timely alerts from automated or manual medicine reminders, and Set reminders based on days, time or frequency. Whether you're trying to get healthier or are simply new to the fitness wagon, Practo keeps you up to date with nifty health & fitness tips that interest you. All tips on Practo are from experienced and certified doctors and health professionals. But in Practo app there is no QR code generation for storing the details about patients, which reduces the paperwork. Users must to carry some files of medical report. Chances of missing file are possible in Practo App. Also it cannot provide facility of hospital recommendation, to easily get nearby hospitals. So that users unable to get proper hospital for particular diseases.

6. PRAPOSED SYSTEM

In proposed system, the classification of diabetes disease is done by SVM, which is binary classifier so, It gives two classes , positive or negative. positive class indicates the patient is diabetes patient and negative class indicates the patient is non-diabetes patient. After successful classification by SVM,K-means clustering used to make cluster of severe, normal and diabetes patients. Then this patients get hospital recommendation depends on the area they belong by using google map. Due to this the patient's appointment is already set at the respective hospital. Patient can carry only the QR code with him/her, which contain the required information about patient.[5]

There are eight numeric variables: (1) Number of times pregnant, (2) Plasma glucose concentration a 2h in an oral glucose tolerance test (3) Diastolic blood pressure (mm Hg) (4) Triceps skin fold thickness (mm) (5) 2-hour serum insulin

(mu U/ml) (6) Body mass index (7) Diabetes pedigree function (8) Age (years). Although the dataset is labeled as there are no missing values, there were some liberally added zeros as missing values. Five patients had a glucose of 0, 28 had a diastolic blood pressure of 0, 11 more had a body mass index of 0, 192 others had skin fold thickness readings of 0, and 140 others had serum insulin levels of 0. After the deletion there were 460 cases with no missing values.

6.1 Training and Test dataset Evaluation

To evaluate the robustness of the SVM models, a 10-fold cross-validation was performed in the training data set. The training data set is first partitioned into 10 equal-sized subsets. Each subset was used as a test data set for a model trained on all cases and an equal number of non-cases randomly selected from the remaining nine datasets. This cross-validation process was repeated 10 times, and each subset serve once as the test data set. Test data sets assess the performance of the models [2].

6.2 Pima Indian Diabetes Dataset

The Pima Indian Diabetes data set [8] was selected from a larger data set held by the National Institutes of Diabetes and Digestive and Kidney Diseases.[10]

There are eight clinical findings (features):

1. Number of times pregnant
2. Plasma glucose concentration a 2 h in an oral glucose tolerance 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
7. Diabetes pedigree function

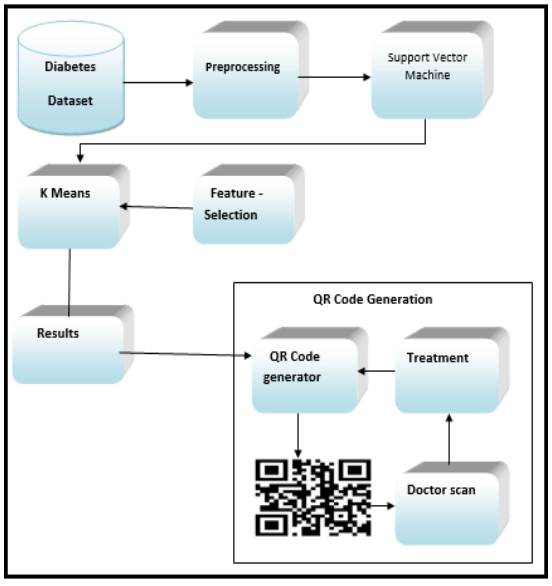

Fig. 2. Architecture of proposed system.

6.3 Quick Response Code

A QR code stands for Quick Response Code is a type of 2D bar code that is used to provide easy access to information through a smartphone. In this process, known as mobile tagging, the smartphone's owner points the phone at a QR code and opens a barcode reader app which works in conjunction with the phone's camera. The reader interprets the code, which typically contains a call to action such as an invitation to download a mobile application, a link to view a video or an SMS message inviting the viewer to respond to a poll. The phone's owner can choose to act upon the call to action or click cancel and ignore the invitation [11].

Fig. 3. Quick Response Code

Static QR codes, the most common type, are used to disseminate information to the general public. They are often displayed in advertising materials in the environment (such as billboards and posters), on television and in newspapers and magazines. The code's creator can track information about the number of times a code was scanned and its associated action taken, along with the times of scans and the operating system of the devices that scanned it. Dynamic QR codes (sometimes referred to as unique QR codes) offer more functionality. The owner can edit the code at any time and can target a specific individual for personalized marketing. Such codes can track more specific information, including the scanners names and email address, how many times they scanned the code and, in conjunction with tracking codes on a website, conversion rates. The technology for QR codes was developed by Densa-Wave, a Toyota subsidiary. The codes were originally used for tracking inventory.QR Code Data capacity Numeric only Max. 7,089 characters Alphanumeric Max. 4,296 characters Binary (8 bits) Max. 2,953 bytes.

7. RESULT ANALYSIS

To analyze the performance of classification, the accuracy and AUC measures are adopted. Four cases are considered as the result of classifier. TP (True Positive) : the number of examples correctly classified to that class. TN (True Negative): the number of examples correctly rejected from that class. FP (False Positive): the number of examples incorrectly rejected from that class. FN (False Negative): the number of examples incorrectly classified to that class. The classification experiments are conducted on the Diabetes dataset. The SVM classifier with RBF kernel is used for classification. The dataset consist of 768 records, which is divided in 80%-20% and used for training and testing respectively i.e. 614 records for training and 154 for testing. In second case it is divided as 60%-40% i.e. we send 461 records for training and 307 records for testing [2].

Algorithm	Accuracy	Sensitivity	Specificity
Decision Tree	75.3%	65%	81%
Naïve Bayes	74.6%	67%	8%
Our SVM	90.2%	72.2%	100%

Fig. 4. Performance Evaluation of Algorithm (80-20)

Algorithm	Accuracy	Sensitivity	Specificity
Decision Tree	74.5%	37.3%	94.5%
Naïve Bayes	77.5%	63.5%	85%
Our SVM	89%	68.2%	100%

Fig.5. Performance Evaluation of Algorithm (60-40)

8. CONCLUSION

The Paper focuses on reducing the patient hard work of caring all the hardcopies of his medical reports. The SVM classifier is used to classify the patient into diabetic positive or negative class then we apply the k-means clustering algorithm to form clusters of diabetic positive patients into normal, moderate and severe diabetic patient. Also the QR code is generated using which the patient need not worry for his reports and it also helps the doctor to better understand the patient medical history. Thus would help the society by using the IT trends like Machine Learning, Web applications and mobile computing.

9. REFERENCES

[1] Khyati K. Gandhi1, Prof. Nilesh B.Prajapati2, "Diabetes prediction using feature selection and classification" International Journal of Advance Engineering and Research Development (IJAERD) Volume 1,Issue 5,May 2014, e-ISSN: 2348 - 4470 , print-ISSN:2348-6406.

[2] Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[3] V. Anuja Kumari1, R.Chitra2," Classification Of Diabetes Disease Using Support Vector Machine" V. Anuja Kumari, R.Chitra / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 2, March -April 2013, pp.1797-1801.

[4] Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[5] Madan Somvanshi, Pranjali Chavan, Shital Tambade, Swati Shinde "A Review Of Machine Learning Techniques using Decision Tree and Support Vector Machine".

[6] S. V. Shinde and U. V. Kulkarni, "Mining Classification Rules from Modified Fuzzy Min-Max Neural Network for data with mixed journal, attributes". Elsevier Journal -Applied Soft Computing pp. 364-378, Dec. 2016.

[7] S. V. Shinde and U. V. Kulkarni, Extended Fuzzy Hyperline-Segment Neural Network with Classification Rule Extraction, Article in Press, Neurocomputing, April 2017.

[8] Clustering - K-means. Available online at: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

[9] Introduction to K-means Clustering. https://www.datascience.com/blog/introduction-to-k-means-clustering-algorithm-learn-data

[10] Vikramaditya Jakkula, "Tutorial on Support Vector Machine" ,2013

[11] Lan Li, Shaobin Ma, Yun Zhang, "Optimization Algorithm based on Support Vector Machine"in Seventh International Symposium on Computational Intelligence and Design, 2014

[12] Burges B.~Scholkopf, editor, "Advances in Kernel Methods--Support Vector Learning", MIT press, 1998.

[13] Patel Brijain R, Kaushik K Rana, "A Survey on Decision Tree Algorithm for Classification", International Journal of Engineering Development and Research (IJEDR), Vol.2, Issue 1, pp.1-5, March 2014