ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A Medical Decision Support System based on Genetic Algorithm and Least Square Support Vector Machine for Diabetes Disease Diagnosis

Aishwarya S[*1], Anto S[2]
[*1] PG Scholar, [2]Assistant Professor, Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, Tamilnadu, India
aish152@gmail.com

Abstract

Clinical decision making is a complex task for physicians since it requires the utmost accuracy of diagnosis. This paper proposes a medical decision support system based on Genetic Algorithm and Least Square Support Vector Machine (LS-SVM) for the diagnosis of diabetes on a Pima Indian Diabetes dataset of UCI machine learning repository. The proposed system uses Genetic algorithm selects a more significant feature subset from the given feature set of the dataset and uses Least Square Support Vector Machine (LS-SVM) for The results show the classification accuracy of the proposed system outperforms that of various existing systems. The performance of the proposed system is analyzed using various parameters like classification accuracy, using 10-fold cross-validation and confusion matrix.

Keywords: Least Square Support Vector Machine, Genetic Algorithm, Diabetes diagnosis, Classification Accuracy.

Introduction

Diabetes is one of the major health challenges which is highly complex and complicated to diagnose for the past few decades. It is caused due to the improper production of insulin in the human body. Insulin is the key parameter responsible to regulate glucose. It leads to many other risks, including kidney disease, blindness, heart disease and never damages. Diagnosis of Diabetes can be done through routine blood checkup. Diabetes can be controlled by proper food habits and exercise program in order to reduce the given risks. Still there is no permanent cure for Diabetes. Diagnosis of diabetes needs special effort for any physician with prior knowledge of the symptoms and deep analysis of the patient's history. Thus to make the diagnosis easier and faster, many machine learning techniques are designed for the automatic diagnosis of Diabetes. Artificial Intelligence (AI), also known as Synthetic Intelligence, is a branch of engineering associated with the computational behavior or intelligent behavior. Essential part of AI in recent years is to simulate human intelligence. Machine Learning is a branch of AI which aims in providing knowledge to such intelligent systems. Machine learning consists of a huge number of algorithms to design and analyze any kind of datasets. Machine Learning can be either supervised or unsupervised. In Supervised learning, data are trained and predicted based on the training. Here, the function is created based on training samples and test on unknown samples. In unsupervised learning, system remains untrained. Decision support systems in the field of medical diagnosis have increased in recent decades. Design of medical expert systems has created more interest among researchers all over the world. Medical expert systems use machine learning techniques for the prediction of any disease based on their existence. Pattern recognition and data mining provide useful retrieval of medical data with the combination of such techniques. Most common data mining technique for decision making from real world data is classification. Usage of data directly could affect the system performance. Features or the attributes have much influence on the performance. Selection of best features will have more impact on the accuracy of the diagnosis system in prediction.

Related Work

Fayssal et al [8] has designed a diagnosis system for diabetes using fuzzy classifier and modified Artificial Bee Colony algorithm. Still the Accuracy of this system is inferior and has paved the way for further research to increase the accuracy. Cheng-Lung Huang et al [9], proposed a general adaptive optimization search methodology and Grid Algorithm combined with SVM classifier. Several real-world datasets such as Diabetes, Heart disease, breast cancer, Contraceptive were validated using the Genetic Algorithm based approach and the Grid algorithm. Average AUC for datasets with two classes using Grid algorithm Diabetes with 0.7647. Hasan et al [10] have used Levenberg–Marquardt (LM) algorithm for training a multilayer neural network structure in order to diagnose diabetes. The obtained accuracy was 82.37%, which is low compared to other existing diabetes diagnosis systems. Yuan ren [11] have proposed two SVM parameter optimization approaches, i.e. GA-SVM and PSO-SVM, adopt an objective function which is based on the leave-one-out cross-validation, and the SVM parameters are optimized by using GA (genetic algorithm) and PSO (particle swarm optimization) respectively. Our proposed system uses the Pima Indian Diabetes (PID) datasets of UCI machine learning repository [8]. The feature selection using Genetic Algorithm is explained in Section 4. Section 5 is dedicated to the classification of PID dataset using Least Square Support Vector machine. The performance of the system on diabetes dataset and other medical datasets are analyzed in Section 6.

Dataset

Pima Indian diabetes data set was obtained from the UCI Repository of Machine Learning Databases [12]. There are 268 (34.9%) cases in class ‘1’ and 500 (65.1%) cases in class ‘0’, Where ‘1’ means a positive test for diabetes and ‘0’ is a negative test for diabetes [9].

Diabetes Attribute information is given below:

1. Number of times pregnant
2. Plasma glucose concentration at 2 hours 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 (weight in kg/ (height in m) ^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

Feature Subset Selection

Selecting appropriate features are important for most data mining methods. Elimination of unimportant features assists in better classification. Usage of the complete and correct dataset will make the system more reliable. The dataset is normalized within a numerical range. The proposed system is shown in Figure.1.

Scaling

Pima Indian Diabetes dataset [12] consists of attributes with different range of values. Usage of these values directly could affect the system stability. In order to avoid any complications in computation the numerical values are linearly transformed to a fixed range using min-max normalization method. The values of the Features from the dataset are normalized between the ranges 0 to 1. The formula used for scaling is given in equation (1).

$$D_{normalized} = \frac{D - D_{min}}{D_{max} - D_{min}}(upperbound - lowerbound) \quad (1)$$

where D is the original data, D_{max} is the maximum value of D, D_{min} be the minimum value in D and $D_{normalized}$ be the normalized valued within the given upper and lower bound. Once scaling is done on the whole dataset feature selection is done to find the most discriminant features. Every feature provides some useful information that could reduce the accuracy of the classifier when the training data is less [5-7]. In order to extract such optimal features Genetic Algorithm is used.

A. Genetic Algorithm

Genetic Algorithm is an evolutionary algorithm which offers multi criterion optimization for higher dimensional space problems. It’s a popular stochastic search method used for feature selection. It is based on Darwin’s theory of natural selection and ‘survival of the fittest’ [4]. Genetic algorithm search initially starts with the least number of attributes. Every set of individuals are called population and each individuals are called as chromosomes. These chromosomes are constituted of many genes which are most binary value indicating the presence of the element in the set. The search of the best result is based on the objective function called as Fitness Function.

Fitness function can be calculated using the formula (2)

$$Fitness = \frac{\text{Total No of Correctly Classified Intances}}{\text{Total No of Training Samples}}$$

(2)

The selected solutions with highest fitness value have more influence than that of the new solutions with less fitness value. This function plays a key role in the selection of the best solution of the problem. In Genetic algorithm, each iteration is known as generation. Fittest individuals are selected from each generation and pooled out to form base for new populations. A new population is created based on the compliance to the fitness function. Offsprings are generated based on the genetic operator's crossover and mutation. Threshold for fitness function will be the maximum accuracy at which the system converges. This process continues till the Fitness threshold is met.

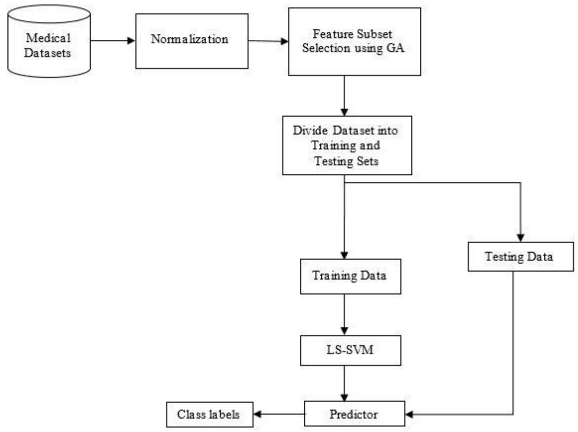


Figure 1: Proposed System

Least Square Support Vector Machine

Least Square Support Vector machine (LS-SVM) is a kind of Support vector machine based on the structural risk minimization principle of statistical learning theory [13]. Support vector machine (SVM) was introduced by Boser [1] and has been used successfully in most regression and classification problems. The key role of SVM in Classification problem is to divide the data into two distinct classes with maximum margin and minimum classification error rate. In order to solve constrained quadratic programming problems, SVM requires higher computational load, which is a major drawback in using in high dimensional problems. In order to overcome this problem, LS-SVM was introduced by Suykens and Vandewalle [2], which uses linear equations to solve the problems.

LS-SVM are used for classification problems to find a hyper plane, which could separate various classes with higher margin. An optimal hyper plane is obtained using maximum Euclidean distance to the nearest point. It maps the input vector into higher dimensional space for non-separable data. Then the optimal separating hyper plane is found.

X_i Є R^p and Y_i Є {0, 1} where X_i is 'p' dimensional input vector and Y_i is the corresponding class label.

$$f(X) = \mathrm{sign}\left(\sum_{i=1}^{N} Y_i \alpha_i K(X, X') + b\right) \qquad (3)$$

Where f(X) be the output of new input vector X_i Є R^p (Equation 3).

X_i is the support vectors belongs to training set. The dataset used for training the classifier are training set.

α_i is the Lagrange multipliers and b be the real constant.LS-SVM performance depends mainly on two key parameters. Choosing best value for these parameters is important to maintain the classifier characteristic. The two kernel parameters are C and Gamma (γ). C be the box constraint and Gamma be the regularization factor[14].

Input data sets are distributed in nonlinear dimensional space. These are converted into high dimensional linear feature space by using kernels. Radial Basis Kernel is used for such mapping for our medical datasets, which is given in (4).

RBF kernels:

$$K(x, x') = \exp\left(-\frac{||x - x'||^2}{\sigma^2}\right)$$

(4)

Performance Evaluation

The proposed System is examined by 10 fold cross validation methodology. The performance of the system is evaluated using four measures: Confusion Matrix, Sensitivity, Specificity and Classification Accuracy.

Confusion Matrix

Confusion matrix [3] (COM) is a 2×2 matrix which shows the predicted and actual classification given in Table 1.

Table 1 Confusion Matrix

Predicted	**Actual**	
	Positive	*Negative*
Positive	TP (true positive)	FP (false positive)
Negative	FN (false negative)	TN (true negative)

- TN is the *correct* predictions of an instance as *negative*.
- FN is the *incorrect* predictions of an instance as *positive*.
- FP is the *incorrect* of predictions of an instance as *negative*.
- TP is the *correct* predictions of an instance as *positive*.

Classification Accuracy

Performance of classifier is commonly measured using classification accuracy (CA). It provides the rate of correctly predicted instances to the overall instances in the dataset. CA can be calculated from Confusion Matrix [3] using the equation (5).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (5)$$

Sensitivity and Specificity

Sensitivity is the true positive rate of prediction, and specificity is the true negative rate [3]. They are defined as in (6) and (7)

$$\text{Sensitivity} = \frac{TP}{TP{+}FN} \qquad (6)$$

$$\text{Specificit y} = \frac{TN}{TN{+}FP} \qquad (7)$$

Experimental Setup

The Proposed System for the diagnosis of diabetes disease is divided into two stages as shown in figure. In the first Stage the Feature selection on the disease dataset is done to reduce the feature space dimension and at this stage different sets of features are obtained. In the second stage, LSSVM classifier is used to classify these feature subsets and the classification accuracy is evaluated. The fittest set of feature subsets with the best classifier parameters are chosen to get an optimal system. The range of C is fixed from 2^{-5} to 2^{15} and Gamma 2^{-10} to 2^{5}.

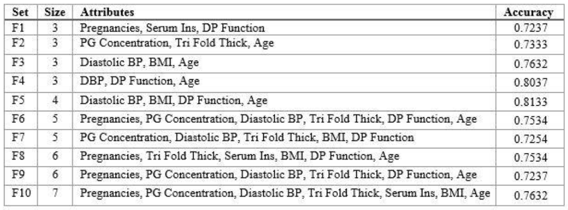
The Process is carried out as below:

Step 1. Different feature subsets are obtained by Feature selection using Genetic Algorithm.
Step 2. Pima Indian Dataset is randomly divided into 10 fold of equal size using k fold cross validation methodology. This is done to maintain the class distribution in each and every fold in the same dataset.
Step 3. First feature subset is fed into LSSVM to get its Fitness value.
Step 4. LSSVM parameters are initialized within the selected range.
Step 5. Classification is performed by using 10fold cross validation
Step 6. Classification accuracy in each fold are calculated and the overall accuracy is obtained.
Step 7: Repeat Steps 3 to 6 for all feature subsets.
Step 8: The feature subset with the highest overall classification Accuracy is chosen as the best discriminating subset.

A. Simulation Results

The diabetes dataset from UCI repository consists of attributes with different numerical ranges and is a complete data set. There are no missing values in this dataset. The numerical ranges are made to be constrained within a fixed range of 0 to 1 using scaling concept. These scaled datasets are fed for feature selection using Genetic Algorithm. The genetic algorithm generates random sequences of subset combinations and uses the Fitness value to predict the fittest subset of features. Here, we have generated 10 Feature Subset from which one optimal subset can be obtained. Table 2 shows the list of all subsets generated. This fitness value depends on the classification accuracy of the system. Classification of these subsets is done by using LSSVM classifier as shown in Table 2. The values of the C and γ parameters should be suitably selected by the user.

Table 2 Feature Subsets using Genetic Algorithm and their accuracy



Set	Size	Attributes	Accuracy
F1	3	Pregnancies, Serum Ins, DP Function	0.7237
F2	3	PG Concentration, Tri Fold Thick, Age	0.7333
F3	3	Diastolic BP, BMI, Age	0.7632
F4	3	DBP, DP Function, Age	0.8037
F5	4	Diastolic BP, BMI, DP Function, Age	0.8133
F6	5	Pregnancies, PG Concentration, Diastolic BP, Tri Fold Thick, DP Function, Age	0.7534
F7	5	PG Concentration, Diastolic BP, Tri Fold Thick, BMI, DP Function	0.7254
F8	6	Pregnancies, Tri Fold Thick, Serum Ins, BMI, DP Function, Age	0.7534
F9	6	Pregnancies, PG Concentration, Diastolic BP, Tri Fold Thick, DP Function, Age	0.7237
F10	7	Pregnancies, PG Concentration, Diastolic BP, Tri Fold Thick, Serum Ins, BMI, Age	0.7632

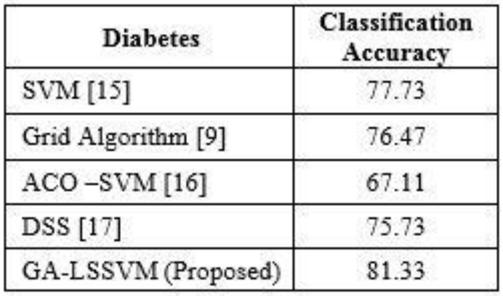
A. Discussions

The proposed system shows a higher performance with feature subsets at an accuracy of 81.33%. The subset consists of Diastolic BP, BMI, DP Function and Age shows the highest fitness and selected as the optimal feature set. The LSSVM parameter γ is chosen to be 0.1. The feature subsets and its accuracy are shown in Table 2. Overall classification accuracy of the proposed system is computed by averaging the classification accuracies of tenfold, which is 81.33%.

Sensitivity and specificity rates of the proposed expert system are obtained as 83.86% and 79.00%, respectively. Classification accuracies of the studies in the literature and our proposed expert system are given in Table 3 for comparison. Performances of all methods given in Table 3 were evaluated on the same Pima Indian Diabetes dataset taken from the UCI machine learning repository.

ISSN: 2277-9655
Impact Factor: 1.852

Table 3: Comparison with existing system



Diabetes	Classification Accuracy
SVM [15]	77.73
Grid Algorithm [9]	76.47
ACO –SVM [16]	67.11
DSS [17]	75.73
GA-LSSVM (Proposed)	81.33

Conclusion

In this paper, a decision support system based on GA-LSSVM is proposed for the diagnosis of the diabetes disease. A Gaussian radial basis function is used as a kernel of LS-SVM. The robustness of the proposed system were analyzed with metrics like classification accuracy, using 10-fold cross-validation and confusion matrix. The accuracy of the system for the PID dataset was found to be 81.33% with GA as a feature selection method. In future, this system can be used for the diagnosis of real life medical data of patients.

References

[1] *B. E. Boser, I. M. Guyon, and V. N. Vapnik, "Training algorithm for optimal margin classifiers," in Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pp. 144–152, Pittsburgh, Pa, USA, July 1992.*

[2] *J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," Neural Processing Letters, vol. 9, no. 3, pp. 293–300, 1999.*

[3] *R. Kohavi and F. Provost, "Glossary of terms," Machine Learning, vol. 30, pp. 271–274, 1998.*

[4] *D.E. Goldberg, Genetic Algorithm in Search, Optimization, and Machine Learning, Addison-Wesley, Boston, 1989.*

[5] *G. V. Trunk," A problem of Dimensionality: A Simple Example", IEEE Trans. Pattern Anal. Mach. Intelligence, vol. 1, pp. 306-307, 1979.*

[6] *A. K. Jain and R. Dubes, "Feature definition in pattern recognition with small sample size", Pattern Recognition, vol. 10, pp. 85-97, 1978.*

[7] *F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection. In: Pattern Recognition in Practice IV, Multiple Paradigms", Comparative Studies and Hybrid Systems, Elsevier, 1994. pp. 403-413.*

[8] *B.Fayssal, M.A.Chikh,"Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm", Computer methods and programs in biomedicine, No.1, pp.92-103, 2013.*

[9] *L.H.Cheng, J.W.Chieh,"A GA - Based Feature Selection and Parameters Optimization for Support Vector Machines" Expert Systems with Applications, Elsevier, Vol.31, pp.231–240, 2006.*

[10] *Hasan Temurtas, Nejat Yumusak, Feyzullah Temurtas,"A comparative study on diabetes disease diagnosis using neural networks", Expert Systems with Applications, Elsevier,Vol.36,pp.8610–8615,2009.*

[11] *R.Yuan, B.Guangchen,"Determination of Optimal SVM Parameters by Using Genetic Algorithm/Particle Swarm Optimization", Journal of Computers, No.5, pp.1160-116, 2010.*

[12] *M.Forina,"Pima Indian Diabetes Dataset", http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes. 1991.*

[13] *Ersen, Y,"An Expert System Based on Fisher Score and LS-SVM for Cardiac Arrhythmia Diagnosis", Computational and Mathematical Methods in Medicine, pp.1-6, 2013.*

[14] *Duygu, C. and Esin,D,"A New Intelligent Hepatitis Diagnosis System: PCA–LSSVM",Expert Systems With Applications,2011.*

[15] *K.C. Tan, E.J. Teoh,Q. Yua,b, K.C. Goh,"A hybrid evolutionary algorithm for attribute selection in data mining", Expert Systems with Applications,Elsevier,Vol 36,2009.*

[16] *Cheng-lung huang ,"ACO-based hybrid classification system with feature subset selection and model parameters optimization",neurocomputing, elsevier, 2009.*

[17] *Massimo Esposito, Ivanoe De Falco∗, Giuseppe De Pietro,"An evolutionary-fuzzy DSS for assessing health status in multiple sclerosis disease",International journal of medical informatics 80,pp.245-254,2011.*