An Expert System based on SVM and Hybrid GA-SA Optimization for Hepatitis Diagnosis

S. Anto, S. Chandramathi

Abstract— An accurate diagnosis of diseases like hepatitis is a challenging task for physicians. This problem in diagnosis has attracted researchers to design medical expert systems with utmost accuracy. This paper proposes a clinical decision support system based on Support Vector Machine (SVM) and hybrid Genetic Algorithm (GA) – Simulated Annealing (SA) for the diagnosis of hepatitis by using the dataset of UCI machine learning repository. The SVM with Gaussian Radial Basis Function (RBF) kernel performs the classification process. The hybrid GA-SA is used for two purposes, one is to select the most significant feature subset of the dataset, and the other is to optimize the kernel parameters of SVM. The performance of the expert system is analyzed using various parameters like classification accuracy, sensitivity and specificity. The classification accuracy of the proposed system is found to be superior to that of the other existing systems in the literature.


1 INTRODUCTION

Machine learning focuses on the improvement of machine projects that develop and change when given new information. It is of two types: Supervised learning and Unsupervised learning. Supervised learning is the task of machine learning that infers a function from labeled training data. Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. There are many examples of machine learning problems such as Optical Character Recognition (OCR), face detection, medical diagnosis and weather prediction. For decision making in data mining, classification is the most common technique used. They help users in understanding how the category relates to other categories.

Utilization of the classification system in medical diagnosis had drawn more attention. Assessment of information taken from patients and decisions of experts are the foundations of diagnosis. Determination of best features has more effect on the precision of the analysis framework in prediction. An automatic diagnosis problem can be approached via both single and hybrid machine learning methods. The proposed work diagnoses hepatitis disease.

Viral hepatitis is a significant health issue around the globe [1]. Hepatitis is the aggravation and damage to hepatocytes in the liver and can be caused by autoimmunity, viruses, infections with fungi and bacteria, or exposure to toxins such as alcohol. Hepatitis diseases can be spread through blood transfusion and shared syringes.

2 RELATED WORKS

Seera et al [2] have proposed a hybrid intelligent system for medical data classification. A hybrid intelligent system that consists of the Fuzzy Min-Max neural network, the classification and regression tree, and the random forest model is proposed. The system yields good generalization performance but is abysmally slow in test phase. The accuracy of the proposed system is 78.39%.

Calisir et al [3] have proposed an intelligent hepatitis diagnosis system using Principal Component Analysis (PCA) and Least Square Support Vector Machine (LSSVM). The simplest invariance could not be captured, unless the training data explicitly provides this information. Kaya et al [4] have proposed a hybrid decision support system based on Rough Set (RS) and Extreme Learning Machine (ELM) for the diagnosis of hepatitis. Redundant features are removed from the dataset through RS approach and classification process is implemented through ELM by using the remaining features. The performance of the proposed system is decreased due to time delay.

Ruxandra Stoean et al [5] have proposed medical decision making model using SVMs, explained by rules of Evolutionary Algorithms (EA) with feature selection. SVMs successfully achieve high prediction accuracy due to a kernel-based engine and EAs can greatly accomplish a good explanation of how the diagnosis was reached. Similarly SVM concept is used for various datasets like diabetes, breast cancer and hepatitis as in literatures [13], [14], [15], [19], [20].

This paper proposes a medical decision support system which uses SVM for classification and GA-SA for optimization as shown in Fig.1. SVM is a kernel based statistical classification technique which is widely used to solve bi-class problems. In this work the SVM uses Gaussian kernel function. GA is an evolutionary algorithm which offers multi criterion optimization for higher dimensional pace problems. It is a popular local search method used for
optimization. Simulated Annealing (SA) is a local heuristic approach which uses greedy technique to solve optimization problems. It is based on the analogy between the simulation of the annealing of solids and the problem of solving large combinatorial optimization problems. In this work the hybridized GA-SA perform optimization to find both the most significant feature sub set and the kernel parameters of SVM.

3 Datasets Used
The proposed method uses hepatitis data set, which is available in the UCI machine learning repository.

In the hepatitis disease dataset, the presence or absence of hepatitis disease is predicted by using the results of various medical tests carried out on a patient. Hepatitis dataset contains 155 instances belonging to two different classes, ‘die’ with 32 instances and ‘live’ with 123 instances. The total number of attributes is 19.

4 Proposed Scheme
The overall process of the proposed system is depicted in Fig.1 which contains Data preprocessing, feature selection, classification and optimization.

4.1 Data Preprocessing
Data preprocessing is a data mining strategy that transforms raw data into a clear format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain numerous errors. By using mean algorithm, the incomplete data is filled.

4.2 Feature Selection
In feature selection, a subset of the original variables is discovered with mean to reduce and dispose the noise dimension. The principle behind feature selection is to pick a subset of input variables by taking out features with little or no predictive information. It transforms high-dimensional data into lower dimensions. When the input to an algorithm is too large to be processed and is suspected to be extremely redundant (much data, but not much information), then the input data is transformed into a reduced representation set of features.

Dimensionality reduction can be achieved by either taking out data that is nearly related to other data in the set, or joining data to make a small set of features [6]. In machine learning, the issue of supervised classification is concerned with utilizing labeled samples to impel a model that classifies objects into a limited set of known classes. The samples are described by a vector of numeric or nominal features. Some of these features may be unimportant or redundant. Avoiding unimportant or redundant features is essential because that they may have a negative impact on the accuracy of the classifier.

Selection of the most significant subset of features is an optimization problem. In this system, the feature selection is done using hybrid GA-SA local search mechanism which is explained in section 4.4.

4.3 Classification
Classification is a data mining function that assigns items in a collection to target classes. The objective of classification is to accurately predict the target class for each case in the data. Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two sets: one for building the model (training) and the other for testing the model. SVM is a class of machine learning algorithms that can perform pattern
recognition and regression based on the theory of statistical learning and the principle of structural risk minimization [9].

Vladimir Vapnik invented SVM for searching a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin [10]. The margin is defined by the distance of the hyperplane to the nearest of the positive and negative examples [11].

Given the training sample of instance-label pairs \((x_i, y_i)\), \(i = 1, \ldots , l\), SVMs require the solution of the following (primal) problem [12]:

\[
\begin{align*}
\min_{w,b,\varepsilon} & \quad \frac{1}{2} w^T W + C \sum_{i=1}^{l} \varepsilon_i \\
\text{Subject to} & \quad y_i (w^T Z_i + b) \geq 1 - \varepsilon_i, \\
& \quad \varepsilon_i \geq 0, i = 1, \ldots , l,
\end{align*}
\]

where the training vector \(x_i\) is mapped onto a high dimension space by mapping function \(\varphi\) as \(z_i = \varphi(x_i)\). \(C > 0\) is the penalty parameter of the error term.

Usually, equation (1) is solved by solving the following dual problem:

\[
\begin{align*}
\min_{\alpha} & \quad F(\alpha) = \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\
\text{Subject to} & \quad 0 \leq \alpha_i \leq C, i = 1, \ldots , l, \\
& \quad y^T \alpha = 0,
\end{align*}
\]

where ‘\(e\)’ is the vector of all 1’s and ‘\(Q\)’ is an \(l\) by positive semi definite matrix.

The \((i, j)^{th}\) element of \(Q\) is given by \(Q_{ij} \equiv y_i y_j K(x_i, x_j)\).

The kernel function is given by equation 3
\[
K(x_i, x_j) = \varphi^T(x_i) \varphi(x_j)
\]

\((\alpha_i)_{i=1}^l\) is Lagrange multipliers, and \(W = \sum_{i=1}^{l} \alpha_i y_i \varphi(x_i)\) is the weight vector.

The classification decision function is,
\[
\text{sgn}(W^T \varphi(x) + b) = \text{sgn}(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b) \quad (4)
\]

The kernel function \(K(x_i, x_j)\) has manifold forms. In this paper, the Gaussian Kernel Function is used. Gaussian kernel function is expressed as follows:
\[
K(x, x_i) = \exp(-\frac{1}{2\sigma^2} ||x - x_i||^2), \quad (5)
\]

(or)
\[
K(x, x_i) = \exp\left(-\frac{1}{\alpha^2} ||x - x_i||^2\right) \quad (6)
\]

Both the equations (5) and (6), which are in the same context, can transform parameter \(\gamma\) and parameter \(\sigma^2\).

The parameters of SVMs with Gaussian RBF kernel refer to the pair : the error penalty parameter ‘\(C\’\) and the Gaussian kernel parameter \(\gamma\), usually depicted as \((C, \gamma)\).

The Gaussian kernel parameter \(\gamma\) is determined by \(\gamma = \frac{1}{\sigma^2}\).

4.4 Optimization

The performance of a SVM classifier depends mainly on the values of the kernel function parameter, Gamma (\(\gamma\)) and penalty function parameter(C). Finding the best values of these two parameters, in order to achieve maximum classification accuracy of the SVM classifier, is
a optimization problem. A hybrid GA-SA algorithm is used to solve this problem and to find the optimal values of C and Gamma.

### 4.4.1 Genetic Algorithm (GA)

GA is a model of machine learning which derives its behavior from a metaphor of some of the mechanisms of evolution in nature. The individuals represent candidate solutions to the optimization problem being solved. In GAs, the individuals are typically represented by n-bit binary vectors [7]. The resulting search space corresponds to an n-dimensional boolean space. It is assumed that the quality of each candidate solution can be evaluated using a fitness function. GAs use some form of fitness-dependent probabilistic selection of individuals from the current population to produce individuals for the next generation. The fitness function is used to search the best result. The selected individuals are submitted to the action of genetic operators to obtain new individuals that constitute the next generation.

Mutation and crossover are two of the most commonly used operators that are used with GAs to represent individuals as binary strings. Mutation works on a single string and generally changes a bit at random while crossover operates on two parent strings to deliver two off springs. Other genetic representations require the use of appropriate genetic operators. The process of fitness-dependent selection and application of genetic operators to generate successive generations of individuals is repeated many times until a satisfactory solution is found.

- **Initial population:** Like in any step by step optimization drawback, the data of great starting parameter advantages the convergence speed of the algorithm. But such kind of data is rarely accessible. This leads to the coverage of large parts of the solution space due to the generation of the random initial population. Hence, to make the exploration of the solution space easier, a heterogeneous initial population is suitable.
- **Selection Step:** According to the fitness value, the individuals are positioned and most elevated rank is given to the best one. At that point, one solution is kept in the next generation.
- **Crossover Step:** The objective of this step is to assemble interesting features of a few solutions in new individuals by making combination of already held solutions.
- **Mutation Step:** This step brings the vital risk for effectively explore the solution space. Any purpose of this space can be reached, it is guaranteed. In addition, if a local optimum is obtained, then a too quick convergence to this neighborhood optimum will be kept away by mutation. Initially, it is set to a maximum value, then decreases to permit convergence and it expands again to avoid local optima.

**Algorithm Steps:**

**Step 1:** Randomly generate an initial source population of ‘n’ chromosomes.

**Step 2:** Calculate the fitness function \( f(x) \) of all chromosomes in the source population using equation (7).

\[
\min f(x) = 100\ n(1.1^2 - x(2))^2 + (1 - x(1))^2 
\]

**Step 3:** Create an empty successor population and then repeat the following steps until ‘n’ chromosomes have been created.

**Step 4:** Using fitness, select two chromosomes \( x_1 \) and \( x_2 \) from the source population.

**Step 4.1:** Apply crossover to \( x_1 \) and \( x_2 \) to obtain a child chromosome ‘n’.

**Step 4.2:** By applying mutation to ‘n’, it deviant new offspring.

**Step 4.3:** Place new offspring in a new population.

**Step 4.4:** Replace the source population with the successor population.

**Step 5:** If the termination criterion is satisfied, stop the algorithm; otherwise go to Step 2.

### 4.4.2 Simulated Annealing (SA)

Annealing is a procedure in metallurgy, in which metals are gradually cooled to make them achieve low energy where they are very strong. In general, SA is a methodology where the temperature is reduced gradually, beginning from a random search at high temperature eventually getting pure greedy descent as it approaches zero temperature.

In optimization, the randomness should be appropriate to jump out of local minima and find regions that have a low heuristic worth. It is a probabilistic meta-heuristic for the global optimization issue of placing a good estimate to the global optimum of a given capacity in a large search space. It is regularly utilized when the search space is discrete [8].

**Acceptance Function**

The calculation chooses which answers to accept so we can better understand how it's able to avoid these local optimums.

Initially, the neighbor solution is checked to see whether it improves the current solution. If so, the current solution is accepted. Else, the neighbour solution is not better and hence the following couple of factors are considered.

- How greatly worse the neighbor solution is?
- How high the current temperature of our system is?

At high temperatures, the system is more likely to accept solutions that are worse.

**Algorithm Steps:**

**Terminology:**

- \( X \) = Design Vector
- \( f_c \) = System Energy (i.e. Objective Function Value)
- \( t \) = Temperature
- \( \Delta \) = Difference in system energy between two
configuration vectors

**Simulated Annealing Algorithm:**

Step 1: Choose a random \( X_i \) select the initial temperature \( t \) and specify the cooling schedule.
Step 2: Evaluate \( f(X_i) \) using a simulation model.
Step 3: Perturb \( X_i \) to get a neighboring Design Vector \( (X_{i+1}) \).
Step 4: Evaluate \( f(X_{i+1}) \) using a simulation model.
Step 5: If \( f(X_{i+1}) < f(X_i) \), \( X_{i+1} \) is the new current solution.
Step 6: If \( f(X_{i+1}) > f(X_i) \), then accept \( X_{i+1} \) as the new current solution with a probability using equation (8)
\[
\exp(-\Delta/t) \quad \text{where} \quad \Delta = f(X_{i+1}) - f(X_i)
\]
Step 7: Reduce the system temperature according to the cooling schedule.
Step 8: Terminate the algorithm.

**Neighborhood Search (Hybrid GA –SA)**

In the hybrid SA algorithm, the best obtained solution in each GA generation is transferred to SA in order to improve the quality of solution through neighborhood search to produce a solution close to the current solution in the search space, by randomly choosing one gene in a chromosome, removing it from its original position and inserting it at another random position in the same chromosome. According to this criterion, even when the value of the next solution is worse, the solution can be accepted based on current temperature to avoid algorithm to stick in a local optimum. In the cooling phase the new temperature is determined by the decrement function \( t \).

**5 EXPERIMENTATION AND RESULTS**

**5.1 Performance Metrics**

To evaluate the prediction performance of SVM classifier, the classification accuracy, sensitivity and specificity are defined and computed.

Confusion Matrix: Confusion matrix shows classifications and predicted. A confusion matrix for a classification problem with two classes is of size \( 2 \times 2 \), as given in Table 1.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>Positive</td>
<td>TP(True positive)</td>
</tr>
<tr>
<td>Negative</td>
<td>FN(False negative)</td>
</tr>
</tbody>
</table>

- TP represents an instance, which is actually positive and predicted by the model as positive.
- FN represents an instance, which is actually positive but predicted by the model as negative.
- TN represents an instance, which is actually negative and predicted by the model as negative.

**Accuracy, Sensitivity and Specificity:**

Accuracy = \( \frac{TP+TN}{TP+FP+FN+TN} \times 100\% \) (9)

Sensitivity is the true positive rate and specificity is the true negative rate. They are defined as follows:-

Sensitivity = \( \frac{TP}{TP+FN} \times 100\% \) (10)

Specificity = \( \frac{TN}{FP+TN} \times 100\% \) (11)

**5.2 Simulation Results**

The system was implemented using MATLAB 7. In this paper, hepatitis disease is diagnosed by using MATLAB software. The hepatitis disease dataset, UCI machine learning repository, is used to validate the proposed system with 80 % tuples for training and the remaining 20 % for testing.

In GA, an average of 30 generations is taken. The best fitness value found is 0.1481 and the mean fitness values for the dataset are also calculated as 0.18 for hepatitis. The creation of generation and convergence of GA for all the dataset- hepatitis is given in Fig.2 and Fig. 3.

For SA, the initial temperature is set as 6.455 and the final temperature as 0.333. The temperature of SA is gradually reduced from initial value to the final in 50 cycles. The stopping criterion for the algorithm is given in equation (8). The GA receives the best chromosome with the help of SA. The working of SA using non-convex function for the medical dataset- hepatitis is shown in Fig.3. The values of \( C \) and \( \gamma \) achieved is \( C=400 \) and \( \gamma = 0.0524 \).

In the hepatitis disease dataset, the total number of attributes is 19, out of which 10 attributes are selected. The selected attributes list is shown in the Table 2.

<table>
<thead>
<tr>
<th>TABLE 2 Selected features of Hepatitis Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SELECTED ATTRIBUTES</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
TABLE 3
ACCURACY, SPECIFICITY AND SENSITIVITY OF THE PROPOSED SYSTEM

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GASA-SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy (%)</td>
</tr>
<tr>
<td>Hepatitis</td>
<td>87</td>
</tr>
</tbody>
</table>

TABLE 4
COMPARISON OF ACCURACY BETWEEN THE PROPOSED METHOD AND THE OTHER METHODS FOR HEPATITIS

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy%</td>
<td>86.12</td>
<td>88.08</td>
<td>74</td>
<td>75.0</td>
<td>83.60</td>
<td>87.0</td>
</tr>
</tbody>
</table>

6 RESULTS AND DISCUSSION

Using 80-20 train-test combination, the proposed system is constructed and tested. The hybrid GA-SA optimization module is used to perform feature selection. The SVM classifier kernel parameters are also optimized using the GA-SA algorithm. The optimized kernel parameters ‘C’ and gamma ensure the best performance of the classifier. The performance of the system is measured using the metrics like classification accuracy, sensitivity and specificity. The details of the performance metrics obtained for the hepatitis dataset is listed in table 3. The comparison of the performance of the proposed system, in terms of the classification accuracy, with the existing systems is given in table 4.

7 CONCLUSION

An expert system based on GASA-SVM is proposed for the diagnosis of hepatitis disease. The hybrid GA-SA is used for the selection of the most significant feature subset of the dataset and also this module optimizes the kernel parameters of the SVM. Achieving the optimal values of C and Gamma leads to the improved classification accuracy of SVM. The performance of the system is tested using various performance metrics like accuracy, sensitivity and specificity. The classification accuracy of the proposed system is compared with that of existing systems.

REFERENCES


