Cost Effective Approach on Feature Selection using Genetic Algorithms and LS-SVM Classifier

E.P.Ephzibah
VIT University, Vellore-632 014
Tamil Nadu, India

ABSTRACT
This work focuses on the problem of diagnosing the disease in the earlier stage by applying a selection technique based on genetic algorithm and least square support vector machines. The implementation of the technique analyses the accuracy of the classifier as well as the cost effectiveness in the implementation. This technique will help us to diagnose the disease with a limited number of tests that could be performed with minimal amount. We use evolutionary computation which is a subfield of artificial intelligence or computational intelligence that involves combinatorial optimization problems. Evolutionary computation uses iterative progress, such as growth or development in a population. This population is then selected in a guided random search using parallel processing to achieve the desired end. Such processes are often inspired by biological mechanisms of evolution. The obtained results using the genetic algorithms approach show that the proposed method is able to find an appropriate feature subset and SVM classifier achieves better results than other methods.

Keywords: Feature selection, Genetic Algorithm, Simulated Annealing, Least Square Support Vector Machines, classification.

1. INTRODUCTION
Data mining consists of a set of concepts and techniques used to find useful patterns within a set of data. The general goal of data mining is to discover knowledge that is not only accurate but also comprehensible and useful for the society. For several years the diagnosis of heart disease had been a complex pattern recognition task. Computer based recognition and classification can achieve high accuracy. Considering the cost for the diagnosis is proved to be helpful for the patients. Diagnosis cost gets reduced if limited number of tests is done instead of all the tests. Datasets that are used for heart diseases involve different features. Some of them are based on laboratory experiments while involve clinical symptoms. However, one of the most popular and useful databases is the UCI Machine learning knowledge repository [2]. There are about 75 data values totally. For the experimental purpose we have taken into consideration the important 13 data values. The knowledge extraction techniques have made it possible to transform various kinds of raw data into high level knowledge. The genetic algorithm is one such knowledge extraction technique which is capable of finding the best subset of features among the other features. It is a general optimization search methodology based on a direct analogy to Darwinian principle of 'survival of the fittest'. Genetic Algorithms work with a set of candidate solutions called a population and optimal solution after a series of iterative computations.

2. EVOLUTIONARY COMPUTATION

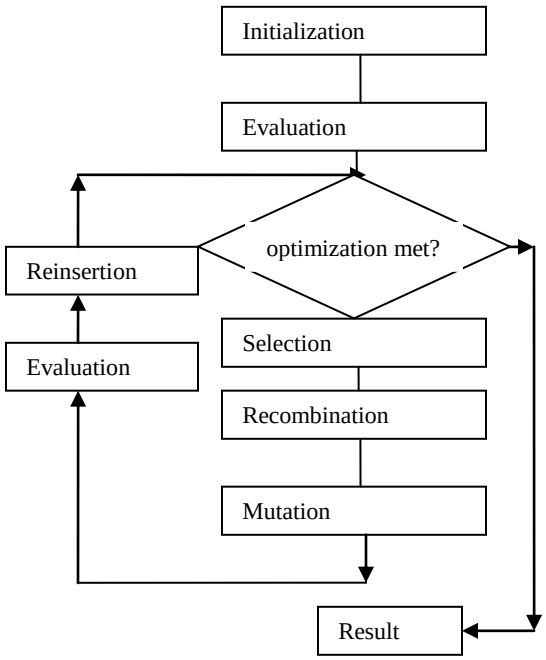

Figure 1: Overview of evolutionary algorithms.

Evolutionary computation is a search technique that is based on the process of natural genetics Darwin's theory of biological evolution. Instead of taking a single solution, evolutionary computation work in parallel on a number of potential solutions for a population of individuals [1]. An overview of the evolutionary algorithm is depicted pictorially in figure1 and its detailed description can be viewed in [15] and [16].

2.1 Genetic Algorithms:
The genetic algorithm has been popularly applied as a tool for optimization problems. This optimization method attempts to incorporate the ideas of natural evolution. It implements a search technique where it constantly tries various possible solutions with genetic operators like selection, crossover and mutation. It is an intelligent exploitation of a random search within a defined search space to solve a problem. There are some steps involved in genetic algorithm which ultimately helps in finding the optimal solution. It is an iterative process.
Genetic Algorithm creates an initial population with a set of solutions. The evaluation process finds the best individual with the help of a fitness function. Sorting the individuals in the population brings out the top best individuals with better fitness

values. The selected individuals stand as the populations for the next generation. Then the genetic operators like crossover and mutation helps in combining the different individuals and changing the feature of an individual. Each solution is called a chromosome. Again the selection operator is applied to bring out the best offspring. The procedure is repeated till the stopping condition is satisfied. In many cases the stopping criteria is the number of generations.

2.1.1 Fitness function:

A fitness function is a particular type of objective function that prescribes the optimality of a solution (that is, a chromosome) in a genetic algorithm so that that particular chromosome may be ranked against all the other chromosomes. Optimal chromosomes, or at least chromosomes which are more optimal, are allowed to breed and mix their datasets by any of several techniques, producing a new generation that will (hopefully) be even better. An ideal fitness function correlates closely with the algorithm's goal, and yet may be computed quickly. Speed of execution is very important, as a typical genetic algorithm must be iterated many, many times in order to produce a usable result for a non-trivial problem.

The fitness function is defined over the genetic representation and measures the quality of the represented solution. The fitness function is always problem dependent. For instance, in the knapsack problem one wants to maximize the total value of objects that can be put in a knapsack of some fixed capacity. A representation of a solution might be an array of bits, where each bit represents a different object, and the value of the bit (0 or 1) represents whether or not the object is in the knapsack. Not every such representation is valid, as the size of objects may exceed the capacity of the knapsack. The *fitness* of the solution is the sum of values of all objects in the knapsack if the representation is valid or 0 otherwise. Once we have the genetic representation and the fitness function defined, GA proceeds to initialize a population of solutions randomly, and then improve it through repetitive application of mutation, crossover, and inversion and selection operators.

2.2 Simulated Annealing:

Simulated annealing (SA) is a probabilistic metaheuristic approach for optimization problems like locating a good approximation to the global optimum of a given function in a large search space. It is based on the analogy between the simulation of the annealing of solids and the problem of solving large combinatorial optimization problems. It is often used when the search space is discrete (e.g., all tours that visit a given set of cities). For certain problems, simulated annealing may be more effective than exhaustive enumeration — provided that the goal is merely to find an acceptably good solution in a fixed amount of time, rather than the best possible solution. With iterative improvement we can derive an approximation algorithm. Once configuration, a cost function and a general mechanism are defined, a combinatorial optimization problem can be solved. Note the acceptance criteria is implemented by drawing random numbers from a uniform distribution on n [0, 1) and comparing these with exp ($-\Delta C_{ij}/c$). [14]

```
PROCEDURE  SIMULATED ANNEALING
Begin
INITIALIZE;
M:=0;
repeat
    repeat
        PERTURB(config.i→ config.j, ΔC_ij);
        if  ΔC_ij <=0 then accept else
        if exp(-ΔC_ij /c) >random [0,1) then accept;
        if accept then UPDATE(configuration j);

    until equillibrium is approached sufficiently
closely;
        C_M+1:=f(C_M);
        M:=M+1;
until stop criterion  = true (System is frozen);
end.
```

Figure 2.Description of the annealing algorithm.

3. FEATURE SELECTION APPROACHES

Feature selection is a process in which it selects a subset of original features. The efficiency of any system can be measured by its classification accuracy. As the dimensionality of a domain expands, the number of features increases. Finding an optimal subset of features is proved to be NP-hard [3]. A problem is NP-hard if solving it in polynomial time would make it possible to solve all problems in class NP in polynomial time. Taking any kind of diagnosis process not all the original features can always be beneficial for classification or regression tasks. There can any many features that are irrelevant or noisy in distribution of dataset. These features are capable of decreasing the classification performance. The feature selection process is necessary for the classification or regression problems in order to obtain higher accuracy with reduced subset of features. When the features are reduced it automatically reduces the cost that is required to do the test.

Methods for data reduction in the context of micro array data analysis or basically feature selection algorithms broadly fall into three categories: the filter model [6] [8] [11] [18], the wrapper model [4], [7], [9], [10], and the hybrid model [5], [13], [17]. Filtering method selects a feature based on its marginal contribution without accounting for its interactions with other features. The selection process is separated from the classification process because a classifier is not built. Another approach for feature selection is the wrapper method which conducts a search for a good subset using an induction algorithm. The algorithm runs on the micro array data, usually partitioned into internal learning and external test sets. The feature subset with highest evaluation is chosen as the final set on which to build a classifier. Maximal classification accuracy on a separate test can be obtained using this approach as the feature subset selection is able to couple tightly with the decision mechanism of the classifier.

4. SUPPORT VECTOR MACHINE

SVMs are useful techniques for data classification. In SVM each object is described by a vector X_i of N real numbers (features or descriptors) which correspond to point in a multidimensional space. The objects in the first class (positive) are each assigned a

value of Y_i= +1, those in the second class are Yi=-1. In linearly separable cases the objects can be correctly classified by

$$w.x_i + b \geq +1 \ for \ yi = +1(class1) \quad \text{--(1)}$$

$$w.x_i + b \leq -1 for \ yi = -1(class2) \quad \text{--(2)}$$

Where w is a vector normal to the hyper plane and b is a scalar quantity.

The SVM attempts to find an optimal separating hyper plane with maximum margin solving the following optimization problem:

$$\underset{w,b}{Max} \frac{2}{IIwII} subject\ to\ y_i(w.x_i + b) - 1 \geq 0 \quad \text{--(3)}$$

The above concepts can also be extended to linearly non-separable cases, in which no hyper- plane can be used to perfectly separate sets of points. In this case, we can introduce non-negative slack variables i=1, 2, 3….m. such that

$$w.x_i + b \geq +1 - \xi_i \ for \quad yi = +1 \quad \text{--(4)}$$

$$w.x_i + b \leq -1 + \xi_i \ for \quad y_i = -1 \quad \text{--(5)}$$

The purpose here is to find a hyper plane that provides the minimum number of training errors i.e. to minimize the constraint violation. The equation to be solved becomes:

$$\underset{w,b}{Max} \frac{2}{IIwII} + c\sum_{i=1}^{m} \xi_i subject\ to\ y_i(w.x_i + b) - 1 + \xi_i \geq 0 \quad (6)$$

Where c is a user predetermined penalty parameter. The parameter c has an important impact on the accuracy of the SVM classifier, thus should be chosen carefully.

The non linear (non-) separable cases could be easily transformed to linear cases by projecting the input variable into new dimensional feature space using a kernel function K (X_i,X_j). Several kernel functions including polynomial, radial basis function (RBF) and sigmoid kernel have been suggested. However radial basis function is the most widely used kernel function and it has been performed very well in most cases. Hence we also have used the Radial basis kernel function.

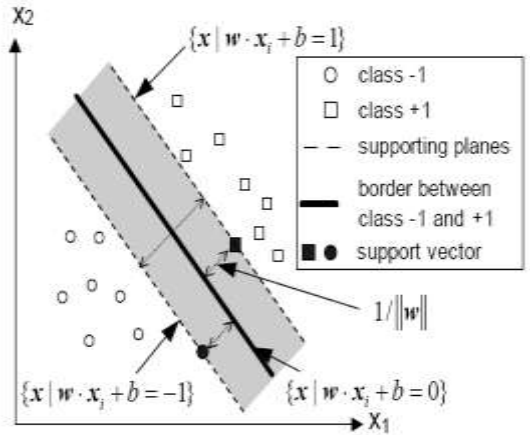

Figure 3: Structure of a simple SVM. [12]

4.1 The least support vector machine classifier:

LS-SVM was proposed by suykens and Vandewalle (1999). In the LS-SVM classification method, the inequality constraints in support vector machines are converted into equality constraints. The training process is done by solving a set of linear equations suykens and Vandewalle (1999). Detailed information on LS-SVM can be found in DTREG (2008). In training and testing of LS-SVM classifier, the choice of the kernel functions has been studied empirically and optimal results have been achieved using Radial Basis Function (RBF) kernel function.

5. DATASET PREPROCESSING

First step that is involved is to refine the data by deleting the chromosomes that has null values. This is to eliminate the bad descriptors and hence reduce the apparent redundancy and overlapping of the descriptors. In this account the following descriptors are removed: (1) descriptors with too many zeros (2) descriptors with very small standard deviation values, and (3) descriptors that are highly correlated with others.

Data scaling: Since values of different descriptors have significantly different numerical ranges, the descriptor values have to be scaled to the same range. In this regard the descriptor values are scaled to the range (-1, +1) by the following formula:

$$V_{scaled} = \frac{v - (1/2)(\max + \min)}{(1/2)(\max - \min)} \quad \text{--(7)}$$

Where V is the original value V_{scaled} is the scaled value, max, min are the maximum and minimum vales of the feature respectively.

6. EXPERIMENTAL WORK

The processed Cleveland Heart Disease database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1, 2, 3, 4) from absence (value 0).

6.1 Attribute Information:

As the work concentrates on the cost that is involved the cost that is required to find the value for each and every attribute take from the UCI ML. As a whole there are 74 attributes out of which 13 attributes have been already selected in the case of Heart Disease dataset. The Fitness functions play a very important role in any genetic algorithm. It decides on which chromosome is selected for further operation. The fitness function is the one which helps us to identify the chromosome which contributes more for the survival of the gene. Hence defining a good fitness function helps us to find out the solution to the problem in a faster and efficient manner.

The gatool had been used for GA implementation from MATLAB R2006b. The selection operator is roulette wheel, and the crossover operator is double one-point crossover, and the mutation operator is binary mutation. The setting of GA parameters is very important in the GA/SVM method. For GA parameters, if the population size is too small, it is difficult to get the best resolution and too big a population size will require a long convergence time. Thus, the size is normally 40–60. If the

crossover *Pc* is too low, it is difficult to search forward and a *Pc* value too big will damage individuals with high adapting value. Therefore, the *Pc* is normally 0.3–0.9. If the mutation rate *Pm* is too low, the new individual is hard to produce and too high a *Pm* would make the GA simple search at random. Thus, the *Pm* is normally 0.01–0.2.

The fitness function in this work emphasizes on the svm accuracy as well as the parameters like Wa and Wf. The values of Wa and Wf are 80% and 20% respectively. The genetic operators like selection, crossover and mutation have the values as "Roulette wheel Selection", 0.8 and 0.05 values respectively. The fitness function (FF) is as given below:

FF: Wa * SVM_accuracy + Wf * Σ (Nf * --(8)

where Wa is the SVM classification accuracy weight, Nf is the number of features selected, costF is the cost of the selected feature, Wf is the weight of the feature number, and SVM_accuracy is the prediction accuracy of the SVM model generated with the given feature subset(Chromosome). costF is an important parameter as the work is based on the cost. The ultimate purpose of the work is to come out with a solution (subset of features) which is cost effective.

LS-SVM classifier	Original set of features	Reduced Feature subset	Accuracy (%)	Cost
Heart Disease Dataset				
without GA	13	-	63	~=600
with GA	13	9 (±2)	83	~=500
Pima Indian Diabetes Dataset				
without GA	8	-	69	~=46
with GA	8	3 (±2)	87	~=20
Breast Cancer Dataset				
without GA	10	-	71	~=100
with GA	10	2 (±1)	86	~<=40

Table 1: Comparison of cost and accuracy with and without Genetic algorithm approach.

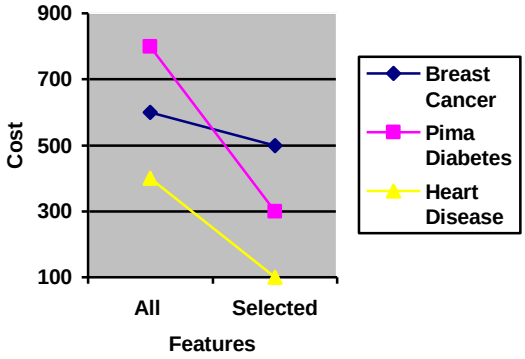

Figure: 4 Cost –Effectiveness graph-Depicting the fall in the cost with GA approach.

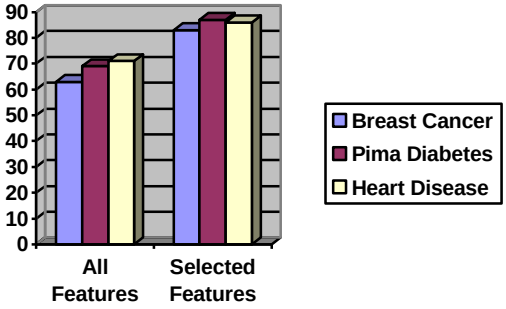

Figure: 5 Graph depicting the Accuracy difference between all features and selected features using GA approach.

7. RESULTS AND CONCLUSION

Genetic programming method was used successfully for investigating machine learning problems in the context of medical classification. Investigation on genetic programming for three of the data collections, such as the Wisconsin breast cancer, the heart disease and the Pima Indian Diabetes data was done. Experiments conducted using GA and LS-SVM classifier prove that this approach comes out with cost effectiveness as well as the performance accuracy.

8. SOCIAL INSIGHT

A good research concentrates on the benefits to the society. Society gets its benefits when the cost effective approaches comes into practice. This research work concentrates mainly on the cost requirement using the evolutionary computing technique called the genetic algorithm approach. Genetic algorithm plays a vital role in finding a feature subset among all the features which can be proved to be more effective.

ACKNOWLEDGEMENTS

Thanks to the authors of the Cleveland Heart Disease Dataset. They would be:
1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.

REFERENCES

[1] Baresel.A: Automating structural tests using evolutionary algorithms,(German) Diploma Theses, Humboldt_University of Berlin, Germany, 2000.

[2] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, vol. 2, pp.121-167, 1998

[3] A.L.Blum and R.L Rivest, "Training a three node Neural Networks is NP-Complete", Neural Networks, vol. 5 , pp.117-127, 1992.

[4] R. Caruana and D. Freitag, "Greedy Attribute Selection", Proc. 11th Int'l Conf. Machine Learning, pp. 28-36, 1994.

[5] S. Das, "Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection," Proc. 18th Int'l Conf. Machine Learning, pp. 74-81, 2001.

[6] M.Dash, K. Choi, P.Scheuermann and H.Liu,
"Feature selection for clustering – a Filter Solution ", Proc. Second Int'l Conference. Data mining, pp.115-122, 2002.

[7] J.G. Dy and C.E. Brodley, "Feature Subset Selection and Order Identification for Unsupervised Learning," Proc. 17th Int'l Conf. Machine Learning, pp. 247-254, 2000.

[8] M.A. Hall, "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning," Proc. 17th Int'l Conf. Machine Learning, pp. 359-366, 2000.

[9] Y. Kim, W. Street, and F. Menczer, "Feature Selection for Unsupervised Learning via Evolutionary Search," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 365-369, 2000.

[10] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, nos. 1-2, pp. 273-324, 1997.

[11] H. Liu and R. Setiono, "A Probabilistic Approach to Feature Selection-A Filter Solution," Proc. 13th Int'l Conf. Machine Learning, pp. 319-327, 1996.

[12] Murphy P M, Aha Irvine D W. CA: University of California, Department of Information and Computer Science[EB/OL].http://www.ics.uci.edu/~mlearn/MLRepository .html,1994.

[13] A.Y. Ng, "On Feature Selection: Learning with Exponentially Many Irrelevant Features as Training Examples," Proc. 15th Int'l Conf. Machine Learning, pp. 404-412, 1998.

[14] P.J.M van Laarhoven and E.H.L.Aarts :Simulated Annealing Theory and applications , (Netherlands : Kluwer Academic Pub-1992),PP 9-10.

[15]Wegener.J. Sthamer, H, Baresel,A (2001): Evolutionary Test Environment for Automatic Structural Testing. Special Issue of Information and Software Technology, vol 43, pp. 851 – 854, 2001.

[16] Wegener.J, Grochtmann ,M: Verifying Timing Constraints of Real-Time Systems by Means of Evolutionary Testing. Real-Time Systems, 15, pp. 275-298, 1998.

[17] E. Xing, M. Jordan, and R. Karp, "Feature Selection for High-Dimensional Genomic Microarray Data," Proc. 15th Int'l Conf. Machine Learning, pp. 601-608, 2001.

[18] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," Proc. 20th Int'l Con Machine Learning, pp. 856-863, 2003.