International Journal of Neural Networks
ISSN: 2249-2763 & E-ISSN: 2249-2771, Volume 2, Issue 1, 2012, pp.-39-43.
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BPJ0000238

COMPREHENSIVE CLASSIFICATION RULES FOR MEDICAL DATA WITH THE INTEGRATION OF GA AND NN

SRIVATHSA P.K.*
Management and Software Consultant, Bangalore, India
*Corresponding Author: Email- pksrivathsa@yahoo.com

Received: October 25, 2012; Accepted: November 06, 2012

Abstract- The present work integrates genetic algorithms (GAs) and NNs to perform the task of classification on the medical data (Bupa Liver data set) that discovers comprehensible IF-THEN rules, in the spirit of data mining. Medical Data mining could be thought of as the search for relationships and patterns within the medical data which facilitates the acquisition of useful knowledge for the effective predictability and diagnosis of diseases. The early detection of any disease certainly facilitates an increased exposure to required patient care with focused treatment, improved cure rates and economic feasibility. The results predict that the methodology is not only reliable but also helps in furthering the scope of the subject.

Keywords- Data Mining, Medical data mining, learning rules.

Citation: Srivathsa P.K. (2012) Comprehensive Classification Rules for Medical Data with the Integration of GA and NN. International Journal of Neural Networks, ISSN: 2249-2763 & E-ISSN: 2249-2771, Volume 2, Issue 1, pp.-39-43.

Introduction

Data mining is an interdisciplinary field that uses methods from several research areas to extract high-level knowledge from real-world datasets. It is an essential part of the broader process: Knowledge Discovery in Databases (KDD) includes several preprocessing methods for preparing data for data mining as well as for post processing methods in order to refine the discovered information. Data mining is increasingly being accepted as a viable means of analysing massive data sets. The semi-automated techniques of data mining could be applied to various domains, but methods from statistics, artificial intelligence, optimization, etc., are not suitable for problems comprising huge data. Generally, the data is noisy and has a high level of uncertainty. Sometimes the data could be dynamic, with the patterns in the 2-D plane of space and time.

Classification task is one of the most studied in the area of data mining. The objective of this task is to predict the value (the class) of a user-specified goal attribute considering the values of the other attributes. Mining classification rules usually utilizes supervised learning techniques that consist in discovering patterns in training data so that the resulting rules can be applied in the classification of other data. To address these aspects of data analysis, heuristic techniques are to be incorporated to complement the existing approaches.

A successful application of Genetic algorithms in machine learning problems is found since 1970s. Genetic Algorithms (GAs) are a search method that has been widely used in applications where the size of the search space is very large. The growth of interest in data mining has motivated the scientific community of evolutionary algorithms. The GAs in classification is an attempt to effectively exploit the large search space associated with it.

The classification task is one of the most studied in data mining which assigns records to one out of a small set of pre-defined classes, by identifying some relationship between attributes. Each record (henceforth an example) consists of a set of predicting attributes and a goal attribute to be predicted [1]. In order to discover the rules detecting some relationship between the predicting attributes and the goal attribute, data mining algorithm is applied to a set of training examples with a known class. The discovered knowledge is usually represented in the form of IF-THEN prediction rules. The evaluation of the discovered rule is subject to the criteria like degree of confidence in the prediction, accuracy rate with regard to classification on unknown-class examples, comprehensibility, etc. In the context of data mining the later happens to be the crucial criterion.

In essence, GAs are “search algorithms based on the mechanics of natural selection and natural genetics” [4]. The principle on which GAs are inspired is the survival of the fittest in the case of fittest individuals selected to produce offspring for the next generation. From the preliminary results reported in this paper it is concluded

that our chromosome encoding and its associated rule set representation are a good alternative for extracting a small set of comprehensible rules and promising. The present GA results seems to be particularly effective in finding a concise set of comprehensible rules, since (by design) it discovers comprehensive rules

Related Work

Some of the works related in this area includes [2-3,6-8]. Very sparse literature in this direction is available. So far no work is available with regard to the present work under investigation.

Medical Data Mining

The idea of medical data mining is to extract hidden knowledge in medical field using data mining techniques. One of the positive aspects of applying data mining techniques to medical data is to extract the hidden knowledge in medical field. It is possible to discover the patterns which are irrelevant. Clinical repositories containing large amounts of biological, clinical, and administrative data are increasingly becoming available as health care systems integrate patient information for research and utilization objectives. Data mining techniques applied on these databases discover relationships and patterns which are helpful in studying the progression and the management of disease [7]. A typical clinical data mining research includes the following structured data, narrative text, data statistics, hypotheses, analysis, interpretation and queries. Prediction or early diagnosis of a disease can be kinds of evaluation. Although data mining is a new technique in the study of medical informatics, it has a rich history in the discovery of patterns. Perhaps it was one of the most successful means of facilitating effective and efficient and decision making in the medical science.

An Overview of GA and NN Approach

Genetic algorithms are the most popular computational models based on evolutionary processes. The four main elements of a genetic algorithm are: (i) genetic code which is a concise representation for an individual solution (ii) population which represents the number of individual solutions (iii) fitness function which represents an evaluation of the usefulness of an individual (iv) propagation techniques which are a set of methods for generating new individuals [Fig-1]. The working procedure of a genetic algorithm is as follows: First, by randomly selecting different genes, a population of individuals is generated. The evaluation of the fitness of each individual is done and to generate a new population (the next generation) the propagation techniques are applied to highly fit the individuals. The cycle of evaluation and propagation continues until the optimal solution is found. Generally, the genetic code is of fixed-length bit string in a genetic algorithm and consequently the population is of fixed size. Elitism, mutation and crossover are the three most common propagation techniques having the following characteristics. The exact individual survives into the next generation in the case of elitism. A new individual is created from an old one in the case of mutation where a small number of randomly selected bits in it's gene is changed, A new individual is created from two old ones in the case of crossover where a random selection of a split point in their genes is done and a creation of a new gene with the left part from one parent and the right part from another is accomplished. The two key aspects of a genetic algorithm are: the genetic representation and the fitness function.

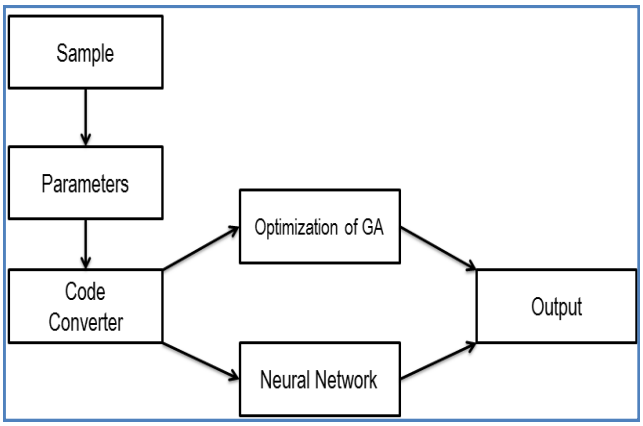

Fig. 1- Block diagram of Genetic Algorithm and Neural network based Classification system

The main objective of genetic feature selection stage is to reduce the dimensionality of the problem before the supervised inductive learning process. Among the many wrapper algorithms used, which solves optimization problems by using the methods of evolution based on the principles of "survival of the fittest" is a promising one. Each individual's fitness as well the quality of the solution are evaluated through GA. In fact, to enter in to the next generation as a population, the fitter individuals have the eligibility and the final set of optimal population with the fitness chromosomes will emerge as the solution after a required number of generations.

Recent researchers are of the opinion that NNs (which are used as learning systems) and GAs (which are used as optimization systems may be combined or integrated in a number of ways resulting in a highly successful adaptive system. If the problem is very complex or has no known solution, the developer is unable to structure the network. In this situation, neural network models include a learning rule which can change the network's structure over the course of training to arrive at the best solution and the most popular learning rule happens to be Back propagation.

Integration of GA and NN

Researchers have combined NNs and GAs in a number of different ways. It is noted that these combinations can be classified into one of two general types (i) Sequential application of NN and GA which is referred to as supportive combinations and (ii) Simultaneous application of NN and GA which is referred to as collaborative combinations [9]. GAs are "search algorithms based on the mechanics of natural selection and natural genetics". The principle on which GAs are inspired is the survival of the fittest which makes a selection of fittest individuals to produce offspring for the next generation. In each generation, the population is evaluated using the fitness function. Next comes the selection process, where in the high fitness chromosomes are used to eliminate low fitness chromosomes. The approached used in the integration of GA and NN are as follows [11]:

Supportive and Collaborative

In a supportive approach, the GA and the NN are applied to two different stages of the problem. Generally, to pre-process the data set that is used to train a NN is GA. In other words, the GA may be used to reduce the dimensionality of the data space by eliminating

redundant or unnecessary features. Since the GA and NN are used very independently supportive combinations are not of at most. In fact, in a collaborative approach, either can be replaced by an alternating technique. It is of interest to note that the GA and NN are integrated into a single system when the population of neural networks is evolved. The main purpose of the system is to find the optimal neural network solution. Certainly such collaborative approaches are feasible since neural network learning and genetic algorithms are complementary approaches. The special features of NN and GA are: (i) A neural network learning rule performs a highly constrained search to optimize the network's structure (ii) A genetic algorithm performs a very general population-based search to find an optimally fit gene.

Evolution of Connection Weights

A genetic algorithm can be applied to optimize a neural network in a variety of ways. According to Yao [10] there are three main approaches viz.,(i) the evolution of weights (ii) the evolution of topology and (iii) the evolution of learning rules.

Evolution of Architectures

In architecture evolution, the genetic code can be either a direct or indirect encoding of the network's topology. However, in the case of direct encoding, each connection is explicitly represented as a binary matrix where 1 indicates the presence of a connection and 0 indicates no connection. While, in an indirect encoding, the important parameters of the network are represented in such a way that the details of the exact connectivity are left to developmental rules (e.g. specify the number of hidden nodes and assume full connectivity between layers). In the above two cases, the specification of the exact neural network is not possible since the weights are determined by the initialization routine and the network's learning algorithm. This results in the evaluation of a noisy gene since it is dependent upon the evaluation of the trained network. Further, GA is capable of finding the best set of architectural parameters rather than the best neural network.

Evolution of Learning Rules

In the final approach, the GA is used similarly to the evolution of architecture. But in the gene the network's learning rule is also encoded as a parametric representation. In this case, the nature of genetic coding of topology is indirect. Neural networks and genetic algorithms are two highly popular areas of research. It is possible to have a highly successful learning systems by integrating both the techniques. Normally, the techniques used are ranged from statistical methods to syntactic and knowledge based approaches. All the above mentioned methods need the definition of a set of features (or symbols and tokens) to be identified and a pattern matcher that is required for the comparison of the observed values which may be ideal and prototypical.

Data Description

We did some experiments with the public domain datasets, in the medical domains of Bupa liver. These data sets were obtained from the UCI (University of California at Irvine) - Machine Learning Repository [5]. These data sets have been used extensively for classification tasks using different paradigms. The data set contains 345 records, each one with 7 attributes such as mean corpuscular volume, alkaline phosphotase, alamine-aminotransferase, aspartate- aminotransferase, glutamyl-transpeptidase, drinks, and selector field used to split data into two sets. The attributes are of nominal one. The first 5 variables are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each line in the Bupa Liver data file constitutes the record of a single male individual. On this database, a sample selector of the drinks>5 is considered.

Experiments and Results

GAssist Intervalar algorithm is used to extract a set of maximally accurate rules that completely defines the feature space.

Hybrid Decision Tree - Genetic Algorithm is used to obtain a Rule Base that better suits the training data by means of a GA search. The Hybrid Decision Tree - Genetic Algorithm (DT-GA) method discovers rules in two training phases. The rule runs C4.5 in the first phase by considering the whole training set and transforms the resulting tree into an "IF-THEN" set of rules. Each rule considers the examples as a small disjunct or as a large disjunct, depending on the criterion: the joint of the instances in the small disjuncts is considered as the second training set in the second phase with GA. The process of pushing down the decision tree continues until it reaches a leaf node and the prediction of an output class of an example will be effective. In the case of a large disjunct leaf node, the class is predicted by the node and the example is assigned to it. In all the other cases, the example is assigned to the class of one of the small-disjunct rules discovered by the GA. The rules from the GA are randomly initialized selecting a value for each attribute from the examples of the "second training set". The encoding of the conditions of all the attributes (nominal and numerical) is in terms of chromosome representation with an additional bit specifying the "don't care" condition. The rule consequent is dynamically chosen as the most frequent class in the set of examples covered by that rule's antecedent but it is not encoded into the genome. Mathematically, the fitness function is given by a quadratic version of the geometric mean associated with the true rates. A specific rule pruning operator is used to transform a condition into a "don't care" type based on the accuracy associated with each attribute. The terminating criterion is given by a threshold value of the remaining examples in the "second training set".

NNEP algorithm is used where a classification model is build means of Product Unit or Sigmoidal Unit Neural Networks to determine a classification model with neural networks based on product unit or sigmoidal unit basis functions. The method consists of obtaining the neural network architecture and simultaneously estimating the weights of the model coefficients with an algorithm of evolutionary computation. In this way a neural network model is obtained from the training set and then checked against the patterns of the generalization set.

Tan_GP is used for mining multiple comprehensible classification rules using genetic programming. an-GP is a GGGP algorithm for classification rule mining which has demonstrated that reports good accuracy and comprehensibility results. It was run using the implementation present in the evolutionary computation framework JCLEC. It is a grammar guided genetic programming algorithm which represents the rules by defining a context-free grammar. The genetic operators considered by this algorithm are crossover, mu-

tation and reproduction. The comprehensible rules are evolved by using: fitness = (tp/(tp+w1*fn)) * (tn / (tn+w2*fp)), where tp: true positives, fp: false positives, tn: true negatives, fn: false negatives and w1 and w2 are weight parameter s respectively.

The implementation of the algorithms is done by using a 10 fold cross validation. [Table-1] presents the accuracies for training and testing data set, sensitivity, specificity for each fold of cross validation.

Table 1- Accuracy Prediction Results Through GAssist Intervalar

Folds	Training				Testing			
	Confusion matrix	Accuracy	Sensitivity	Specificity	Confusion matrix	Accuracy	Sensitivity	Specificity
1	95 35 39 141	76.12	73	78.33	5 10 7 13	51.42	33.33	65
2	97 33 29 151	80	74.61	83.88	8 7 9 11	54.2	53.33	55
3	97 33 25 159	81.5	74.16	86.41	10 5 5 11	67.7	66.66	68.75
4	99 31 24 157	82.3	76.15	86.74	7 8 4 15	64.7	46.66	78.94
5	87 45 19 159	79.3	65.9	89.32	7 6 6 16	65.7	53.84	72.72
6	91 41 28 150	77.7	68.93	84.26	11 2 9 13	68.57	84.61	59
7	90 41 19 160	80.6	68.7	89.38	6 8 6 15	60	42.85	71.42
8	105 25 38 142	79.6	80.76	78.88	8 7 6 14	62.8	53.33	70
9	84 46 24 156	77.4	64.61	86.66	8 7 7 13	60	53.33	65
10	81 49 19 161	78	62.3	89.44	11 4 2 18	82.8	73.33	90

Table 2- Prediction Of Prime Complexities Through Hybrid-decision tree GA

Folds	Evaluation time	Selection time	Cross over	Mutation	Replacement time	Total time	Fitness (Training)	Fitness (Testing)
1	12.76	3.17	3.01	0.032	0	19.78	57.18	21.66
2	15.87	4.15	3.39	0.03	0.015	24.11	62.58	29.33
3	22.88	4.75	3.25	0	0.016	31.64	64	45.82
4	41.21	7.65	5.67	0.2	0.03	56.2	65.84	36.83
5	29.9	4.82	5.02	0.14	0	41.12	58.86	39.15
6	30.01	4.72	5.72	0.1	0.016	41.93	58.08	49.65
7	53.85	9.06	8.15	0.16	0.024	73.04	61.4	30.6
8	41.79	9.03	7.56	0.188	0.16	60.4	63.7	37.33
9	42.55	6.3	5.79	0.22	0.015	57.11	55.99	34.66
10	85.81	11.41	8.41	0.43	0.047	108.48	55.72	65.99

Fitness: 0.6513158563136073

Number of hidden neurons: 4 Number of effective link: 16

Train CCR: 76.12903225806451

Test CCR: 77.14285714285714

Alpha Input: 0.033456391882986776

Alpha Output: 0.06691278376597355

Success Ratio: 0.205474018224672748

Fig. 2- NNEP Algorithm Results

Table 3- Parameter Description for Tan_GP Algorithm

Parameter Descriptor	Value
Population-size	150
Max-generations	100
Max-deriv-size	20
Rec-probability	0.8
Mutation-probability	0.1
Copy-probability	0.01
W1	0.7
W2	0.8
Elitist-probability	0.06
Support	0.03

We have used 10 fold cross validation to implement the algorithms. [Table-1] gives the accuracies for training and testing data set, sensitivity, specificity for each fold of cross validation. [Table-2] gives the and fitness function for testing and training samples, evaluation, selection, crossover, mutation, replacement and total time for each fold. [Table-3] gives the parameter description for Tan_GP Algorithm

[Table-4] gives the accuracy prediction for each fold for the training and testing samples. [Table-5] gives the rules generated for diagnosis of the liver disease. [Fig-2] gives results for NNEP algorithm.

Table 4- Accuracy Prediction Results Through Tan_GP

Folds	Training Accuracy	Testing Accuracy
1	58.38	60
2	70.6	51.4
3	69.42	51.6
4	63.9	58.8
5	55.8	48.5
6	61.6	57.1
7	58.7	57.1
8	65.8	54.2
9	67.4	57.1
10	68	60

Table 5- Rules Generated For Diagnosis

Folds	Rules	Training	Testing
1	IF gammagt <= 156.0 THEN Class = 1 IF mcv > 86.0 AND gammagt > 27.0 THEN Class = 2 IF alkphos <= 86.0 AND gammagt > 13.0 THEN Class = 2 IF gammagt > 8.0 AND drinks > 0.5 THEN Class = 2 IF sgot > 20.0 AND gammagt > 12.0 THEN Class = 2 IF mcv > 79.0 AND drinks <= 6.0 THEN Class = 1 (108/264)	80	65.71
2	IF sgpt > 16.0 AND sgot > 15.0 THEN Class = 2 (140/246) IF sgpt <= 86.0 AND gammagt > 11.0 THEN Class = 1 IF mcv > 87.0 AND drinks <= 8.0 THEN Class = 2 IF drinks <= 10.0 THEN Class = 1 (124/305) IF alkphos <= 101.0 AND drinks > 6.0 THEN Class = 2 (21/41)	76.77	60
3	IF sgpt > 37.0 AND gammagt > 16.0 THEN Class = 2 (34/57) IF mcv > 82.0 AND sgot > 18.0 THEN Class = 2 (141/235) IF alkphos > 59.0 AND drinks > 0.5 THEN Class = 2 IF gammagt > 5.0 AND drinks <= 7.0 THEN Class = 1 (109/273) IF drinks > 4.0 THEN Class = 2 (48/88)	78.66	74.19
4	IF gammagt <= 52.0 AND drinks > 0.5 THEN Class = 1 (63/152) IF alkphos <= 69.0 AND drinks > 0.0 THEN Class = 2 (103/168) IF mcv <= 97.0 AND alkphos <= 85.0 AND sgpt <= 58.0 AND sgot > 12.0 AND gammagt > 4.0 AND drinks > -0.5 THEN Class = 2 (131/221)	77.49	64.7

Table 5- Continue

5	IF sgpt <= 87.0 AND drinks <= 4.0 THEN Class = 1 (88/214) IF gammagt > 18.0 AND drinks <= 20.0 THEN Class = 2 (127/200) IF sgpt > 17.0 AND gammagt > 10.0 THEN Class = 1 (106/243) IF sgpt > 18.0 AND gammagt <= 14.0 THEN Class = 2 (13/42)	74.83	54.28
6	IF mcv <= 99.0 AND drinks <= 12.0 THEN Class = 1 (128/302) IF sgot > 13.0 AND drinks <= 10.0 THEN Class = 2 (170/288) IF gammagt <= 114.0 AND drinks <= 6.0 THEN Class = 1 (110/261)	74.51	77.14
7	IF sgot <= 50.0 AND drinks > -0.5 THEN Class = 2 (173/303) IF gammagt <= 135.0 AND drinks > -0.5 THEN Class = 1 (129/301) IF mcv <= 98.0 AND drinks > 3.0 THEN Class = 2 (89/136)	74.19	42.85
8	IF sgot <= 34.0 AND drinks <= 4.0 THEN Class = 1 (88/198) IF gammagt > 13.0 THEN Class = 2 (157/254) IF gammagt <= 201.0 AND drinks <= 20.0 THEN Class = 1 (129/305) IF alkphos <= 80.0 AND drinks <= 3.0 THEN Class = 2 (69/131) IF mcv > 83.0 AND gammagt > 5.0 THEN Class = 1 (122/288)	77.41	68.57
9	IF alkphos <= 78.0 AND gammagt <= 84.0 AND drinks <= 4.0 THEN Class = 2 (95/151) IF sgot <= 25.0 AND gammagt <= 66.0 THEN Class = 2 (102/188) IF mcv <= 97.0 AND alkphos > 46.0 AND sgpt > 14.0 AND sgot > 12.0 AND gammagt > 4.0 AND drinks <= 2.0 THEN Class = 2 (61/123) IF gammagt <= 108.0 AND drinks <= 6.0 THEN Class = 1 (104/257)	76.77	68.57
10	IF drinks > 3.0 THEN Class = 1 (52/142) IF sgot <= 24.0 AND gammagt > 14.0 THEN Class = 2 (74/134) IF mcv <= 98.0 AND alkphos > 73.0 AND sgpt <= 148.0 AND sgot > 17.0 AND gammagt <= 159.0 AND drinks > -0.5 THEN Class = 2 (51/90) IF alkphos > 48.0 AND drinks <= 4.0 THEN Class = 2 (113/191)	78.38	82.85

Conclusion

The most common difficulties met in the medical data statistical analysis come from the following facts: we have to deal, in the most cases, with a very large number of parameters, and the parameters are often very different in their nature – the physician has to make a through observation of many parameters that characterize the patient's condition in order to establish an accurate and correct diagnosis. Further, the information has to be retrieved from all the possible sources. It is a well known fact that the physicians experience accumulated during many years of practice results in a correct diagnosis. Although, the knowledge intuition and experience of a good physician cannot be replaced from the computer results still researchers are striving hard in designing the so-called "expert systems" – computerized applications which are able to establish the diagnosis, based on the values of a large amount of parameters collected regarding the patient's condition.

In the present investigation, the classification rules for medical mining (Bupa liver data set) are generated by integrating GA and NN approach. The algorithms used are GAssist Intervalar, Hybrid decision tree-genetic, NNEP algorithms and Tan_GP. The discovered knowledge is represented in the form of IF-THEN prediction rules which are of high level and symbolic knowledge type contributing to the comprehensibility of the discovered knowledge. It is important to note that the discovered rules can be effectively evaluated according to several criteria like degree of confidence in the prediction, the accuracy rate of classification on unknown-class examples, comprehensibility etc., Finally, it is concluded that the results facilitate the early effective detection of the diseases and the survival rate.

References

[1] Abdelaal A.M.M. and Farouq W.M. (2010) *International Multi Conference on Computer Science and Information Technology*, 11-17.

[2] Araujo D.L.A., Lopes H.S. and Freitas A.A. (1999) *Proc. IEEE Systems, Man and Cybernetics Conf.*, Tokyo, 3, 940-945.

[3] Bellaachia A. and Erhan G. (2006) *Ninth Workshop on Mining Scientific and Engineering Datasets in Conjunction with the Sixth SIAM International Conference on Data Mining*.

[4] Chen A.L., Hubbard H., Schatz S.M. (2000) *Artificial Intelligence Review*, 437-466.

[5] Merz C.J. and Murphy P.M. (1996) *UCI Repository of Machine Learning Databases Irvine*, CA.

[6] Peña-Reyes C.A, Sipper M. (2000) *Artificial Intelligence in Medicine,* 19(1), 1-23.

[7] Schaffer D., Whitley D. and Eshelman L. (1992) *International Workshop on Combinations of Genetic Algorithms and Neural Networks*, CA, 1-37.

[8] Srimani P.K. and Manjula S.K. (2011) *AIP Conf.*, 1414, 51-55.

[9] Srivathsa P.K. (2011) *AIP Conf.*, 141, 67-71.

[10] Yao X., Liu Y. (1997) *IEEE Transactions on Neural Network,* 8, 3, 694-713.

[11] Zhijiang J., Shengzhong F. (2009) *Fifth International Conference on Natural Computation,* 71-79.