Hybrid learning of wrapper and embedded method for feature selection of medical data

Aye Mya Thandar
University of Computer Studies, Yangon, Myanmar
ayemyathandar7@gmail.com

Abstract

Several recent machine learning publications demonstrate the utility of using feature selection algorithms in many learning. Feature selection helps to acquire better understanding about the data by telling which the important features are and how they are related with each other and it can be applied to both supervised and unsupervised learning. This paper aims to find the best subset of features that not only maximizes the classification accuracy but minimizes the number of features. The other reason is to make aware of the necessity and benefits of applying feature selection methods. In this paper, genetic algorithm is one of the wrapper feature selection methods and it is used to reduce the irrelevant attributes of data. Embedded feature selection method (C4.5) is used to prune the features selected by genetic algorithm which is suffering from overfitting problem. By combining genetic algorithm with decision tree, this method enhances the Bayesian classification to eliminate unnecessary features and produces accurate classifier.

Keywords: Genetic Algorithm, Decision tree, feature selection, Bayesian Classifier

1. Introduction

Machine learning provides methods, techniques, and tools that can help solving diagnostic and prognostic problems in a variety of medical domains .There are many methods for improving the speed and accuracy of machine learning programs on large data sets, especially those in which the data objects have large numbers of features. Feature selection plays an important role in data selection and preparation for data mining and machine learning [7]. Feature subset selection is the process of identifying and removing as much of the irrelevant and redundant information as possible. This reduces the dimensionality of the data and allows learning algorithms to operate faster and more effectively. In some cases, accuracy or feature classification can be improved because an improvement in accuracy of a fraction of percent might translate into significant savings.

Among feature selection methods, filter method selects a feature subset as preprocessing steps where features are selected based on properties of the data itself and independent of the induction algorithm [12].Filter method is simple, fast and easily scale to very high-dimensional datasets. However, each feature is considered separately depending on ignoring feature dependencies and it lead to worse classification performance. Wrappers approach utilizes the learning algorithms itself as a criterion in selection features. Feature selection and training are carried out in a single step during system design. Therefore, wrapper method has ability to take into account feature dependencies and it also has a higher risk of overfitting than filter techniques.

Some researchers found that using a decision-tree to select features for use in the Bayesian classifier gave good result. The decision tree algorithm usually uses an entropy-based measure known as “information gain” as a heuristic for selecting the attribute that will best split the training data into separate classes. The main advantage of decision tree is that it can provide post-pruning to overcome overfiiting problem of genetic algorithm [5]. However, its disadvantages are that it ignores feature dependencies and its classification capability may be bad. In this case, the training set is classified using genetic algorithm to reduce irrelevant attributes. There exists some redundant attributes which will affect the classification accuracy of medical result and even lead to the wrong decisions. Attribute reduction deletes some irrelevant or unimportant attributes while maintaining the attributes of classification and decision-making ability.

Hybrid approaches are presented to solve the classification problem and feature selection. This approach combines Genetic algorithm and Decision Tree algorithm to search for the useful subsets of features for classifying medical datasets. The aim of the proposed method is to find a subset of relevant attributes that leads to a reduction in both the classification error rate and attributes (features).

2. Related works

It is known that Naïve Bayesian classifier (NB) works very well on some domains, and poorly on some. The performance of NB suffers in domains that involve correlated features. C4.5 decision trees, on the other hand, typically perform better than the Naïve Bayesian algorithm on such domains, [3] it describes a Selective Bayesian classifier (SBC) that simply uses only those

features that C4.5 would use in its decision tree when learning a small example of a training set, a combination of the two different natures of classifiers. Selecting relevant genes from the microarray data poses a formidable challenge to researchers due to the high-dimensionality of features, multi-class categories being involved, and the usually small sample size. To overcome this difficulty, [2] a filter method (information gain, IG) and a wrapper method (genetic algorithm, GA) was proposed for feature selection in microarray data sets. IG was used to select important feature subsets (genes) from all features in the gene expression data, and a GA was employed for actual feature selection. The K-nearest neighbor (K-NN) method with leave-one-out cross-validation (LOOCV) served as an evaluator of the IG-GA.

Decision tree is less used as a credit scoring model because its classification accuracy is easily affected by noise data and the redundancy attribute of data. Therefore, some researchers consider combining decision tree with other data mining techniques. Classification model [5] is built based on a decision tree by learning historical data. Clustering algorithm and genetic algorithm are combined to further improve the accuracy of credit scoring model by removing noise data and reducing redundancy attributes. In this paper, hybrid learning of wrapper and embedded method for feature selection is proposed. Hybrid approaches are presented to solve the classification problem and feature selection. This approach combines Genetic algorithm and Decision Tree algorithm to search for the useful subsets of features for classifying medical datasets.

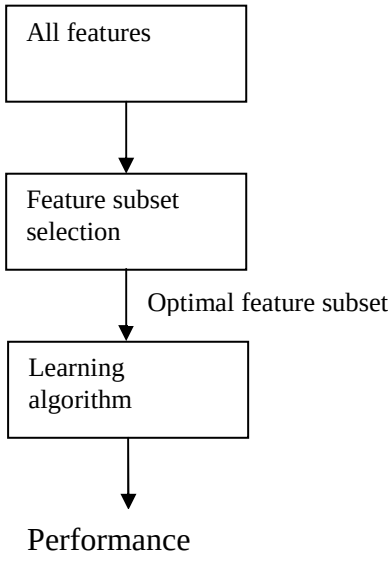
3. Feature Selection methods

Feature selection is the process of removing features from the data set that are irrelevant with respect to the task that is to be performed. Feature selection can be extremely useful in reducing the dimensionality of the data to be processed by the classifier, reducing execution time and improving predictive accuracy [4]. Feature selection algorithms may be widely categorized into three groups: filter, wrapper method and embedded methods.

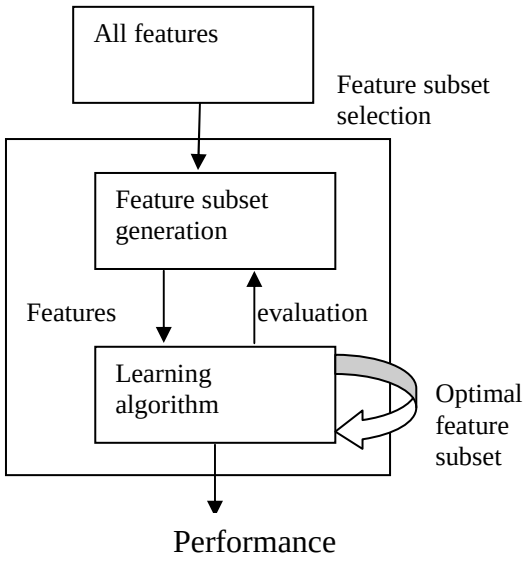
Filter methods do not require the use of a classifier to select the best subset of features. These methods attempt to assess the merits of features from the data, ignoring feature dependencies and the effects of the selected feature subset on the performance of the learning algorithm. However, filter methods are relatively computationally cheap, since they do not involve the induction algorithm.

Wrapper methods conduct a search for a good subset using the learning algorithm itself as part of the evaluation function. Advantages of Wrapper methods include the interaction between feature subset search and model selection, and the ability to take into are that they have a higher risk of over fitting than filter techniques.

The embedded approach embeds the selection within the basic induction algorithm. Comparing with the wrapper model, feature selection algorithms of embedded model are usually more efficient, since they look into the structure of the involved learning model and use its properties to guide feature evaluation and search. In recent years, the embedded model is gaining increasing interests in feature selection research due to its superior performance. Examples of Embedded feature selection methods are Id3, C4.5, 1-norm support vector machine and sparse logistic regression etc [8]



(a) Filter method



(b) Wrapper method

Figure1. Two approaches to feature selection

4. The Proposed Hybrid learning using Genetic Algorithm and Decision Tree

This paper describes a hybrid methodology that integrates genetic algorithm and decision tree learning to find the best subset of features that not only maximizes the classification accuracy but minimizes the number of features. Genetic algorithm (GA) is used to search for the space of all possible subsets of a large set of candidate discrimination features. GA is easy to parallelize and it utilizes the induction algorithm itself as a criterion in selecting features since in the context of learning classification rules. Each of the selected feature subsets is evaluated (its fitness measured) by testing the decision tree.

Searching for the best subset in very large spaces is prone to overfitting, even if assessment relies on cross-validations. This method is first attempt and the advantage of GA becomes more obvious when overfitting problem is solved by post-pruning method of decision tree and it provide good accuracy of optimal feature selection. GA and decision tree algorithm (DT) can complement each other. Another point of this paper is to reduce the weakness of DT by using the advantages of GA. DT is an unstable feature selector and it ignores feature dependencies and the effects of the selected feature subset on the performance of the learning algorithm. With the optimal feature selected by GA, DT can overcome the prospect of feature dependencies problem.

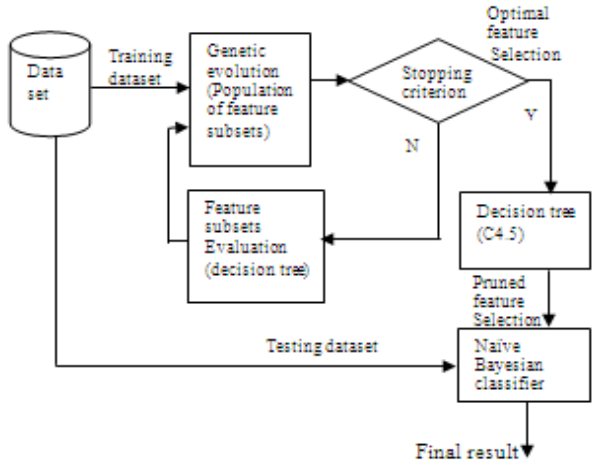


Figure2. Hybrid learning system using genetic algorithm and decision tree

In this hybrid leaning system (figure- 2), medical datasets are used as training data. First, genetic algorithm reduces irrelevant attributes by calculating its fitness function. Accuracy rate of decision tree algorithm (C4.5) on testing data is used as fitness function of genetic algorithm. Stopping criterion is defined by number of generation. After finishing all generations, optimal feature selection is got and reduced attributes are calculated again by decision tree algorithm (C4.5) which is also one of the embedded feature selection methods and it overcomes the prospect of feature dependencies problem. Then, it produces pruned features and these features are classified by Naïve Bayesian classifier to produce final result. This hybrid feature selection method will be shown the better accuracy of medical diagnosis by testing these selected features in Naïve Bayesian classifier. Accuracy and selected feature will show in term of final result

5. Genetic algorithm

Genetic algorithm (GA) is stochastic search algorithm modeled on the process of natural selection underlying biological evolution. It have been successfully applied on a variety of problems including scheduling problems, machine learning problems, multiple objective problems, feature selection problems, data mining problems and traveling salesman problems. GA proceeds in an iterative manner by generating new populations of strings from old ones. Every of string is the encoded binary, real etc., version of a candidate solution.

The parameters used in this system are represented in a string of binary digits, chromosome and each digit is called a gene. GA representation and meaningful fitness evaluation are the keys of the success in GA applications. An evaluation function associates a fitness measure with every string and indicates its fitness for the problem. In this paper, the classification performance of the decision tree (DT) on unseen data is used as a measure of fitness for the given feature sets. Standard GA applies genetic operators such as selection, crossover and mutation on an initially random population in order to compute an entire generation of new strings. This system will use the size of population and number of generation is 20 respectively and stopping criterion is defined by generation. The probability of crossover is 0.6 and mutation probability is 0.003 as optional value.

```
Simple Genetic Algorithm ()
{
      Initial population;
      Evaluate population;

   While termination criterion not
reached
     {
        Select solutions for next
population;
        Perform crossover and mutation;
        Evaluate population;
     }
```

Figure3. Simple genetic algorithm

6. Decision tree algorithm (C4.5)

Decision tree (DT) algorithm is a very popular and efficient data-mining technique. It builds an interpretable model that represents a set of rules and it is relatively fast to train and make predictions. DT is an unstable feature selector and the number of features selected by DT is strongly related to the sample size.

The C4.5 algorithm is a descendent of ID3, which builds decision trees top down and prunes them. The tree is constructed by finding the best single feature test to conduct at the root node of the tree. After the test is chosen, the instances are split according to the test, and the sub problems are solved recursively; C4.5 prunes by using the upper bound of a confidence interval on the re-substitution error as the error estimates, since nodes with fewer instances have a wider confidence interval, they are removed if the difference in error between them and their parents is not significant.

6.1 Information Gain

Attribute relevance analysis is to compute some measure that is used to quantify the relevance of an attribute with respect to a given class and such measures include information gain. This measure was achieved by an independent ranking of each feature using an information theory based information measure to estimate which features are the most discriminatory. Let S be a set of training samples. Suppose the class label attribute has m distinct values defining m distinct classes, C_i. Let s_i be the number of samples of S in class C_i with probability s_i /s, where s is the total number of samples in set S. The expected information needed to classify a given sample is

$$\mathbf{I\ (s_1, s_2, s_3, \ldots\ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2 (p_i)}$$

An attribute A has v distinct values $\{a_1, a_2, \ldots, a_v\}$ can be used to partition S into the subsets$\{S_1, S_2, \ldots, S_v\}$, where Sj contains those samples in S that have value a_j of A. Let Sj contain s_{ij} samples of class C_i. The expected information based on this partitioning by A is known as entropy of A. It is the weighted average:

$$\mathbf{E(A) = \sum_{j=1}^{v} \frac{S1j+S2j+\ldots..+Smj}{S}\ I(S1j,S2j,\ldots,Smj)}$$

The information gain obtained by this partitioning on attribute A is defined by

$$\mathbf{InformationGain\ (A) = I(s_1, s_2, s_3, \ldots, s_m) - E(A)}$$

The attribute with the highest information gain is chosen as the test attribute for the current node. Such approach minimizes the expected number of tests needed to classify an object and guarantees that a simple tree is found.

6.2 Information Gain Ratio

A simple decision tree algorithm only selects one decision tree given an example set, though there may be many different trees consistent with the data. The information gain measure is biased in that it tends to prefer attributes with many values rather than those with few values. C4.5 suppresses this bias by using an alternative measure called Information Gain Ratio, which considers the probability of each attribute value. The Split Information takes into account the factor of an attribute having many values. It is defined as

$$\mathbf{SplitInformation\ (A) = -\sum_{j=1}^{v} \frac{|Sj|}{|S|} \log_2 \frac{|Sj|}{|S|}}$$

And the gain ratio is

$$\mathbf{GainRatio\ (A) = \frac{InformationGain\ (A)}{SplitInformation\ (A)}}$$

6.3 Tree Pruning

C4.5 builds a tree so that most of the training examples are classified correctly. Though this approach is correct when there is no noise, accuracy for unseen data might degrade in cases where there is a lot of noise associated with the training examples and/or the number if training examples is very small. To alleviate this so-called overfitting problem, C4.5 uses the post-pruning method. This approach allows C4.5 to grow a complete decision tree first, and then post-prune the tree. It tries to shorten the tree in order to overcome overfitting. This generally involves removal of some of the nodes or subtrees from the original decision tree. Its goal is to improve the accuracy on the unseen set of examples by pruning.

7. Classification Accuracy

Accuracy is estimated as the number of correct class predictions, divided by the total number of test samples. In this paper, Naïve Bayesian classifier is use to classify the attributes selected by GA and DT. This classifier estimates the probability of the features given the class for each class. This system will use true positive, true negative, false positive and false negative to calculate accuracy. True positive answer denotes correct classifications of positive case (true positive-TP). True negative answer denotes correct classifications of positive case (true negative-TP). False positive answer denotes incorrect classifications of

negative cases into class positive (False positive-FP). False negative answer denotes incorrect classifications of negative cases into class positive (False negative-FP). The classification accuracy measures the proportion of correctly classified cases:

Accuracy = (TP+TN) / (TP+TN+FP+FN)

8. Medical data and experimental evaluation

The system uses 5 datasets from the UCI repository. For example, breast cancer dataset contains 192 instances with 13 attributes and 4 classes. Breast cancer is a leading cause of cancer death among women. Digital mammography is one of the most suitable methods for early detection of breast cancer. The high percentage of unnecessary biopsies are performed and many deaths caused by late detection or misdiagnosis. A computer based feature selection and classification system can help to get better accuracy and result.

Firstly, this system use and run 13 attributes of breast cancer dataset by using genetic algorithm and it selects 4 attributes called lumpsize, pain, invasivesymptoms and signofcarcinoma. Genetic algorithm reduces irrelevant attributes such as age, patientsymptoms and class because breast cancer disease does not depend on age and other information. Then, it also reduces redundant attributes such as lumpposition, lumpduration and natureoflump.

In this paper, the other reason of using genetic algorithm is to solve dependency problem of decision tree algorithm. For example, an older person can get much salary then younger one as age and money are related
with each other .In breast cancer dataset, natureoflump and lumpduration depend on each other. If lumpduration is shorthistory, nature of lump usually can be soft. Therefore, genetic algorithm reduces natureoflump and lumpduration.

According to a doctor's decision, attribute pain is important to determine breast cancer result because when any patient finds her lump does not pain and she may be suffer from breast cancer disease. Therefore, lumpsize measure may be noise data and can cause overfitting problem. Decision tree algorithm (C4.5) is used to solve this problem and it selects pain, invasivesymptom and signofcarcinoma. Selected attributes do not affect the result and accuracy of any dataset.

8.1. Attributes used in breast cancer

Table1.Attributes in breast cancer datasets

Attribute	Value
Age	Real
Recurrence	no-rec,rec
Personaldata	latemarriage,early menstruation, latemenopause, no-child, no-breast-feeding, none
Familyhistory	present,none
Lumpposition	unilateral,bilateral, upperouter, central, remainingregion
Lumpduration	longhistory, shorthistory
Natureoflump	Hard, soft
Lumpsize	<2cm,2-5cm,>5cm
Pain	painless, painful
Patient Symptoms	present, none
Invasivesymptoms	auxiliarynodes,chest, liver, yellowishskin, bone, none
Signsof Carcinoma	elevated,retracted, eccentric, bleeding, dimplinglikeorange, sorebreast, none
Result	I, II, III, IV

8.2 Sample Train Data of breast cancer

'<35','no-rec','latemarriage','present','upper outer',’shorthistory’,’hard','>5cm','painless', 'present','Axillary nodes’, ‘Elevated', 'IV‘

'35-50','rec','no-breast-feeding’,’present','unilateral','longhistory’, ’soft','2-5cm','painless','present', 'none','Eccentric', 'III‘

'35-50','rec','latemarriage','none','central', 'longhistory’, ‘soft','<2cm','painful', 'none','none','Bleeding', 'II‘

'35-50','rec','no-child','present','central','long history’,’ soft','2-5 cm','painful','present', 'none’, ‘none', 'I'

Table2. Dataset detail before feature selection

Dataset	Attribute	Class	Instances
Hepatitis	20	2	155
Sick	30	2	3163
Lung-cancer	57	3	128
Hypothyroid	30	4	3772
Breast cancer	13	4	192

Table3. Dataset detail after feature selection

Dataset	Attribute	Attributes selected by GA	Attributes selected by C4.5
Hepatitis	20	10	6
Sick	30	19	17
Lung-cancer	57	13	5
Hypothyroid	30	15	11
Breast cancer	13	4	3

Table4. Accuracy detail of Bayesian classifier before feature selection of hybrid learning

Dataset	NB (Accuracy) %	NB (Inaccuracy) %
Hepatitis	84.5161	15.4839
Sick	92.6034	7.3966
Lung-cancer	90.625	9.375
Hypothyroid	80.351	19.649
Breast cancer	88.7006	11.2994

9. Conclusion and future work

With the current rapid increase in the amount of biomedical data being collected electronically in critical care and the wide-spread of cheap and reliable computing equipment, many researchers have already started, or eager to start, exploring these data. In spite of the increase in the incidence of the disease, the death rates of breast cancer continue to decline. This decrease is believed to be the result of earlier breast cancer analysis and classification as well as improved treatment. This paper is to predict patient's breast cancer result based on their diagnosis using Naïve Bayesian classifier with the best feature selection of hybrid method and these plans will work well in future. Therefore, this system will use breast cancer dataset as training data first, and test some new diagnosis. And it will also compare the accuracy and number of features in 5 datasets from the UCI repository before and after feature selection based on hybrid of decision tree and genetic algorithm.

References

[1] Alaa. M. Elsayad. (2010) "Predicting the Severity of breast masses with ensemble of Bayesian classifiers", Department of Computers and systems, Electronics Research Institute, Egypt.

[2] Cheng-Huei Yang, Li-Yen Chuang and Cheng-HongYang (2009) "IG-GA: A Hybrid Filter/ Wrapper Method for Feature Selection of Microarray Data", Taiwan

[3] Chotirat Ann Ratanamahatana & Dimitrios Gunopulos. (2004) "Scaling up the Naïve Bayesian Classifier: Using Decision Trees for Feature Selection", University of California.

[4] C.N.Hsu, H. J. Huang and S. Dietrich (2004), "The ANNIGMA-Wrapper Approach to Fast Feature Selection for Neural Nets", IEEE Transactions on System, Man and Cybernetics, Part B, vol. 32, no. 2, pp.207-212.

[5] Defu Zhang, Stephen C.H.Leung, Zhimei Ye,(2008) "A Decision Tree Scoring Model Based on Genetic Algorithm and K-means Algorithm", Third International Conference on Convergence and Hybrid Information Technology.

[6] Duda, R.O., & Hart, P.E. (1973). "Pattern classification and scene analysis". New York, NY: Wiley.

[7] G.John,R. Kohavi, and K. Pfleger,(1994) " Irrelevant Features and the Subset Selection Problems", Proceedings of 11th Int'1 Conference on Machine Learning, San Mateo, CA,pp121-129.

[8] Huan Liu, Hiroshi Motoda, Feature Selection: An Ever Evolving Frontier in Data Mining, JMLR: Workshop and Conference Proceedings 10: 4-13 The Fourth Workshop on Feature Selection in Data Mining ,2010.

[9] Huiquing Lin, Jinyan Li, Limsoon Wong (2004), " A Comparative Study on Feature Selection and Classification Using Gene Expression Profiles and Proteomic Patterns", Laboratories for Information Technology, 21 Heng Mui Keng Terr, 119613 Singapore.

[10] J. Bala,J . Huang and H. Vafaie & K.Dejong and H.Wechsler (1995) " Hybrid Learning Using
Genetic Algorithms and Decision Trees for Pattern Classification",IJCAI conference.

[11] Ron Kohavi, (2000) "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", Data Mining and Visualization Silicon Graphics, Inc, 2011 N.Shoreline Blvd, Mountain View,CA 94043-1389

[12] S.Nirmala Devi and Dr.S.P Rajagopalan(2011). "A study on Feature Selection Techniques in Bio-Informatics", International Journal of Advanced Computer Science and Applications

[13] Genetic Algorithms, A step by step tutorial, Max Moorkamp, Dublin Institute for Advanced Studes, Barcelona, 29th November 2005

[14] Introduction of genetic algorithm for geophysical applications, Gaspar Monsalve (2008)

[15] Feature-selection ability of the decision tree algorithm and the impact of feature-selection/extraction on decision-tree results based on hyper spectral data, International Journal of Remote Sensing archive Volume 29 Issue 10,May 2008.