Scaling up the Naive Bayesian Classifier using Genetic and Decision Tree for feature selection

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

Abstract

This paper aims to scale up the Naïve Bayesian Classifier using Genetic and Decision Tree for feature selection. The main reason is to predict patient's breast cancer result based on their diagnosis using this scaled classifier. Naïve Bayes can suffer from oversensitivity to redundant and/or irrelevant attributes. Several researchers have emphasized on the issue of redundant attributes, as well as advantages of feature selection for the Naïve Bayesian Classifier. In this paper, Genetic algorithm is used to reduce redundant attributes in feature selection, and then apply Decision tree algorithm to find an optimal set of feature weights that improve classification accuracy. By combining genetic algorithm with decision tree, and this method enhance the Bayesian classification to eliminate unnecessary features and produce fast, accurate classifiers. Bayesian classifier represents each class with a probabilistic summary, and finds the most likely class for each example it is asked to classify.

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

1. Introduction

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 can be found in many areas of data mining such as classification, clustering, association rules, regression. Early research efforts mainly focus on feature selection for classification with labeled data (supervised feature selection) where class information is available. Feature extraction aims to reduce the computational cost of feature measurement, increase classifier efficiency, and allow greater classification accuracy based on the process of deriving new features from the original features.

Bayesian classifier is simple, but it will not be optimal when attribute independence does not hold. It is known, Decision Tree typically perform better than the Naïve Bayesian algorithm on such domains. Some researchers found that using a decision-tree to select features for use in the Bayesian classifier gave good result. One of the problems with using decision tree when there are too few training 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. Decision trees examples available is that it might give a constant decision without generating the decision tree. And classification accuracy of Decision Tree is easily affected by noise data and the redundancy attributes of data. Decision Tree algorithm called C4.5 can generate IF-THEN rules. The main advantage of it is that it can provide intelligence classification rules for decision-makers to help them understand the contents of the data sets and make the correct decision. However, its disadvantages are that 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 breast cancer 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.

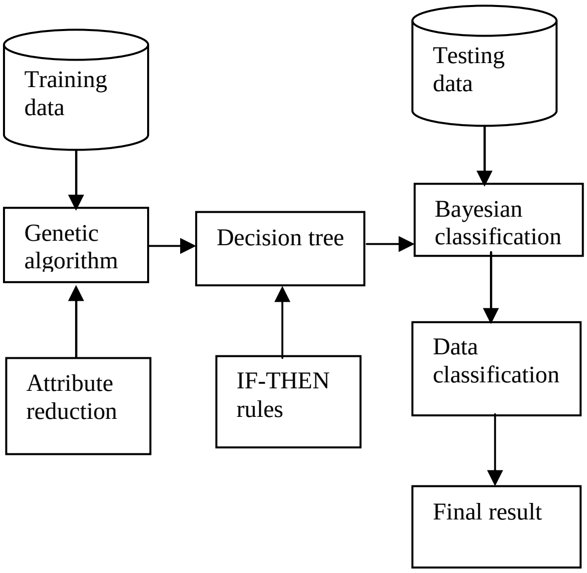
2. Related Works

[1]Cestnik (1990) reached similar conclusions, [6]Kononenko(1990) reported that, in addition, at least one class of users(doctors) finds the Bayesian classifier’s representation quite intuitive and easy to understand, something which is often a significant concern in machine learning. [7]Kubat, Flotzinger, and Pfurtscheller (1993) found that using a decision-tree learner to select features for use in the Bayesian classifier gave good results in the domain of EEG signal classification. [3]The simple Bayesian classifier is limited in expressiveness in that it can only create linear frontiers (Duda & Hart, 1973). Therefore, even with many training

examples and no noise, it does not approach 100% accuracy on some problems.

[8] Langley (1993) proposed the use of “recursive Bayesian classifiers” to address this limitation. In his approach, the instance space is recursively divided into subregions by a hierarchical clustering process, and a Bayesian classifier is induced for each region. Although the algorithm worked on an artificial problem, it did not provide a significant benefit on any natural data sets. In a similar vein, [5] Kohavi (1996) formed decision trees with Bayesian classifiers at the nodes and showed that it tended to outperform either approach alone, especially on large data sets. [9]Ron Kohavi scale up the accuracy of Naive-Bayes Classifiers by using Decision-Tree Hybrid in 2000. [4]That paper has described a new algorithm, NBTree, which is a hybrid approach suitable in learning scenarios when many attributes are likely to be relevant for a classification task. In 2004, three researchers compare the study on feature selection and classify methods using gene expression profiles and proteomic patterns.[2] A hybrid of a Decision Tree scoring model based on Genetic algorithm and K-means algorithm has recently been proposed in 2008.

3. System Overview



3.1. Attributes used in the system

Attribute	Value
Age	Real
Recurrence	no-rec,rec
Personaldata	latemarriage,earlymenstruation, 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
Patientsymptoms	present, none
Invasivesymptoms	auxiliarynodes, chest, liver, yellowishskin, bone, none
Signsofcarcinoma	elevated, retracted, eccentric, bleeding, dimplinglikeorange, sorebreast, none
Result	I, II, III, IV

3.2 Sample Train Data

'<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','no-rec','no-child','none','upperouter’, ‘shorthistory’,’hard','25cm','painless','none', 'Aillary nodes','Elevated', 'IV‘

'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'

3.3 Accuracy comparison

Train size	Bayesian classifier (Accuracy)	Bayesian classifier (Inaccuracy)	Decision tree (Accuracy)	Decision tree (Inaccuracy)
50	88.2353%	11.7647%	97.0588%	2.9412%
100	88.6364%	11.3636%	97.6754	2.3246
200	90.9605%	9.0395%	99.435%	0.567%
300	91.635%	8.365%	99.2395%	0.7605%

4. Classification theorems

4.1 Bayesian Classification

Bayesian approach to classification is to estimate the probability of the features given the class for each class and then use Bayes Rule to get the desired quantity, the probability of the class given the features. The formalize this, let C be the classification, which can take one of several values, C1, C2... Cn. In order to make the classification, a set of features called $x=x_1$, x2,…,xd are used as input patterns for neural networks. These are n possible classes, and d features are used to make the classification. Obviously, if the values of the features express the probability of the classes, good decisions about which classification can be made. For example, the classification which would give the least misclassification errors can be made. P (C|x) is hard to estimate. The approach taken is to estimate the probability of x for each class and then use Bayes rule to invert. Bayes rule is as below

$$P(C|x) = \frac{P(x|C)\,P(C)}{P(x)}$$

The terms on the right hand side are,

P (x|C) This is the likelihood of the features given the class. Since there are few classes, we can use standard techniques of estimating probabilities from data for each class to estimate this.

P(C) This is the prior. One way to get this is the frequency of each class in the data. This is only valid if those frequencies are representative of the frequencies in the real world. If not, you need prior information to estimate the probability of each class.

P(x) this can be computed from the identity,

$$P(x) = \sum_{i=1}^{n} P(x|C_i)\,P(C_i)$$

Where the sum is over all classes.

4.2 Decision Tree Algorithm

Most algorithms that have been developed for learning decision tree are variations on a core algorithm that employs a top-down, greedy search through the space of decision trees. The attribute should be tested at the root of the tree is evaluated using a statistical test to determine how well it alone classify the training examples. The best attribute is selected and used as the test at the root node of the tree. A descendant of the root node is then created for each possible value of this attribute, and the training examples are sorted to the appropriate descendant node. The entire process is then repeated using the training examples associated with each descendant node to select the best attribute to test at that point in the tree. To select the test attribute at each node in the tree the information gain measure is used. The algorithm never backtracks to reconsider earlier choices. The ID3 is a famous algorithm to construct a decision tree. And the C4.5 is the extended version of the ID3. C4.5 is a well-known induction algorithm which uses information-theoretic concepts to grow a decision tree. It first grows a full tree and then retrospectively prunes it in order to avoid overfitting. C4.5 converts this tree to a set of rules which can be further pruned. The paper uses C4.5 to construct a scaled Bayesian classifier for feature selection.

4.3. Genetic Algorithms

Genetic algorithms are optimization algorithms based on some of the processes observed in natural evolution. Genetic algorithms are viewed as domain independent search methods. They are iterative search procedures used to find optimal solutions. But it can, under the right circumstances, find acceptably short period of time. The search is done by having a population of individuals. The individuals are expressions of solutions. Each solution (chromosome) is a combination of bit strings (genes). Three basic genetic operators called selection, crossover and mutation guide to reduce attribute. GAs using cross-validation to evaluate a given feature subset show in some cases a significant overfitting problem. Using leave-one-out error bounds instead is an alternative. It leads to a better generalization performance in most cases, but if the number of features to select is not fixed beforehand, a higher number of features is selected than with cross-validation. Optimizing kernel parameters within the GA is useful, especially if leave-one-out error bounds are used.

5. Conclusion

With the current rapid increase in the amount of biomedical data being collected electronically in critical care and the wide-spread availability 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 this scaling Bayesian classifier. To do this paper, we have to prove how these plans work well in future. Therefore, we will use breast cancer dataset as training data first, and test some new diagnosis. By comparing the accuracy of only using Bayesian classifier and scaling Bayesian classifier based on decision tree and genetic algorithm, we can express how this improved classifier work correctly.

References

[1] Cestnik, B. (1900). Estimating probabilities: A crucial task in machine learning. Proceedings of the Ninth European Conference on Artificial Intelligence. Stockholm, Sweden: Pitman.

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

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

[4] Huiquing Lin, Jinyan Li, Limsoon Wong,A Comparative Study on Feature Selec9tion and Classification Using Gene Expression Profiles and Proteomic Patterns, Laboratories for Information Technology, 21 Heng Mui Keng Terr, 119613 Singapore(2004)

[5] Kohavi, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (pp. 202-207). Portland, OR: AAAI Press.

[6] Kononenko, I. (1990). Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. In B. Wielinga (Ed.), Current Trends in Knowledge Acquisition. Amsterdam, The Netherland: IOS Press.

[7] Kubat, M., Flotzinger, D., &Pfurtscheller, G. (1993). Discovering patterns in EEG-Signals: Comparative study of a few methods. Proceedingd of the Eighth European Conference on Machine Learning (pp.366-371). Vienna, Austria: Springer-Verlag.

[8] Langley, P. (1993). Induction of recursive Bayesian classifier. Proceedings of the Eighth European Conference on Machine Learning (pp. 153-164). Vienna, Austria: Springer-Verlags.

[9] Ron Kohavi, 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(2000)