Int J Pharm Bio Sci 2015 April; 6(2): (B) 182 - 190

ResearchArticle Bioinformatics

International Journal of Pharma and Bio Sciences

ISSN 0975-6299

SIGNIFICANCE OF INFORMATION GAIN RATIO FOR IMPROVING CLASSIFICATION OF HEART DISEASES

R.KARTHIKEYAN[1] A.KUMARAVEL[2] AND V.KHANAA [3]

[1]Assistant Professor, Department of Computer Science and Engineering. Bharath University, Selaiyur, Chennai-600073, India, rkarthikeyan1678@gmail.com
[2]Professor and Dean, Department of Computer Science and Engineering. Bharath University, Selaiyur, Chennai-600073, India,drkumaravel@gmail.com
[3] Professor, and Dean Department of Information Techonology. Bharath University, Selaiyur,

ABSTRACT

Mechanizing the prediction of new patients' heart disease diagnosis based on data mining on historical data is an extremely useful tool in the cardiology stream. There exist many studies focusing on this specific aspect of the filtering the attributes. The objective of this research paper is two-fold. First, we look into four distinct classifiers for evaluating the relevancy of the attributes and we investigate the effects of feature selection in such experiments.

KEY WORDS: **Heart disease data set, Information gain, Decision trees, Decision rules, Meta classifiers, Bayes classifiers, Function classifiers.**

R.KARTHIKEYAN
Assistant Professor, Department of Computer Science and Engineering. Bharath University, Selaiyur, Chennai-600073, India, rkarthikeyan1678@gmail.com

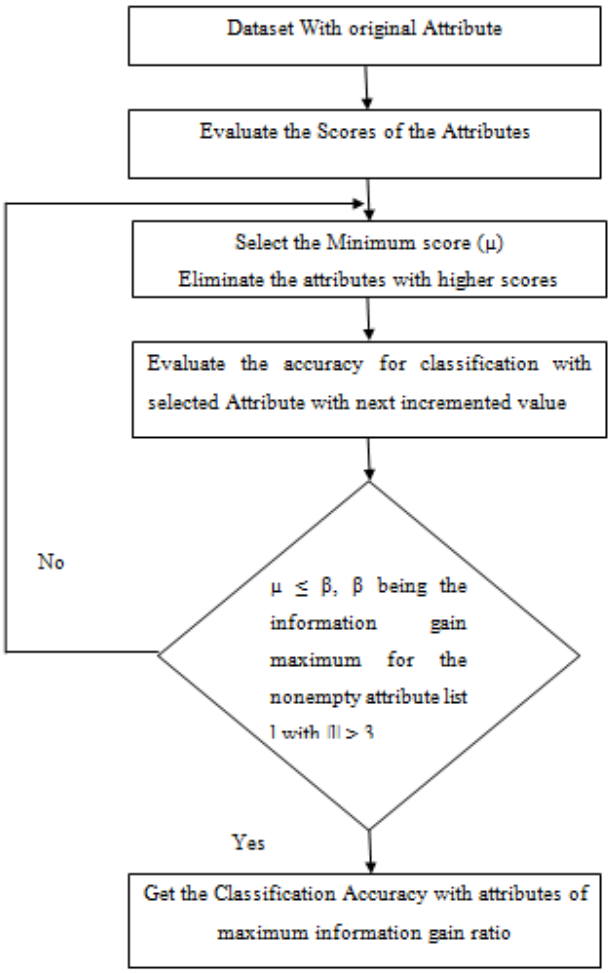
***Corresponding author**

INTRODUCTION

Physicians for diagnosing heart disease search for minimal information to predict the problem class with maximum accuracy. The research work with same orientation are available in [1,2,3].In this paper we perform selection of attributes by an iterative method based on information gain ratio as seen in [4].Attribute selection can be defined as a process that chooses a minimum subset of attributes from the original set of attributes, so that the search space is optimally reduced according to a certain evaluation criterion. Identifying such subset happens to be NP-hard as it involves exponential time complexity with respect to the number of attributes. We organize with data set description in section 2.We describe the main method for attribute selection in section 3 followed by results and description in section 4

Figure 1

Steps for classification with selection of attributes based on Information gain ratio



2. DATA PREPARATION

We present here the details of data collection and the recommended tool for mining such collected data. The data collected is downloaded from UCI data repository from the internet. The primary method along the detailed steps is depicted in the diagrammatic format as in Fig 1.

2.1 Dataset

The datasets has been slightly modified for these experiments in order to make it in 'attribute relation flat file (arff) format'. The data sets are downloaded from [14].

2.1.1 Dataset description

The heart disease datasets describe with a set of fourteen attributes as shown in the list 2.2 below. The class attribute has discrete values {0,1,2,3,4}. The two distinct data sets obtained from two different geographical regions are Cleveland data and Switzerland data.

2.2 List of description of attributes

1. Age (In years)-Numerical
2. Sex (1 = male; 0 = female)-Nominal
3.cp-chest pain type –Nominal
4. trestbps- testing blood pressure (in mm Hg on admission to the hospital)- Numerical
5. chol -serum cholestoral in mg/dl- Numerical
6. fbs -(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)-Nominal
7. restecg- resting electrocardiographic results-Numerical
8. thalach maximum heart rate achieved-Numerical
9. exang -exercise induced angina (1 = yes; 0 = no)- Nominal
10. oldpeak ST depression induced by exercise relative to rest-Numerical
11. slope- the slope of the peak exercise ST segment-Numerical
12. ca number of major vessels (0-3) colored by flourosopy-Nominal
13. tha 3 = normal; 6 = fixed defect; 7 = reversible defect- Nominal
14. class {0,1,2,3,4}-Nominal

Table 1
Statistical Description of attribute

S.No	Attribute Name	Mean	Standard Deviation
1.	Age	54.4	9.0
2.	Sex	0.6	0.5
3.	cp	3.1	1.0
4.	Trestbps	131.6	17.6
5.	Chol	246.6	51.7
6.	Fbs	0.14	0.3
7.	Restecg	1.0	1.0
8.	Thalach	149.6	22.8
9.	Exang	0.3	0.5
10.	Oldpeak	1.0	1.2
11.	Slope	1.6	0.6
12.	Ca	0.6	0.9
13.	Tha	4.7	1.9

2.3 Method

Here the method for ranking the attributes with their 'scores' is considered. Ranker Search Method makes attributes by their individual evaluations Use in conjunction with attribute evaluators with the parameter generate ranking (true or false), number to select, and threshold values is set threshold by which attributes can be discarded. Default value results in no attributes are discarded. Use either this option or number to select to reduce the attribute set. The classification, variable ranking is a filter method: it is a preprocessing step, independent of the choice of the predictor. The ranker method generally performs the rank which attributes should be obtain high or low rank according to the selected attribute in the given datasets. Ranker is providing a rating of the attributes, orderly by their score to the evaluator. Algorithms of ranking rank the attributes in the dataset. The results were validated using different algorithms for classification. A large set of classification algorithms is available, each with its strengths and weaknesses. There is no single learning algorithm that works best on all supervised learning problems. Four widely used supervised learning algorithms are adopted here to build models, namely, IB1, Naive Bayes, C4.5 decision tree and the radial basis function (RBF) network. An advantage of Naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. C4.5 decision tree has Various advantages: simple to understand and interpret, requires little data preparation, robust, performs well with large data in a short time. RBF network offers a number of advantages, including requiring less formal statistical training, ability to implicitly detect complex nonlinear relationships between dependent and independent variables, ability to detect all possible interactions between predictor variables, and the availability of multiple training algorithms.

2.3.1 Main Method

Entropy is commonly used in the information theory measure, which characterizes the purity of an arbitrary collection of examples. It is in the foundation of the IG, GR, and SU attribute ranking methods. The entropy measure is considered a measure of the system's unpredictability. The entropy of Y is $H(Y)=-\sum_{y\in Y} p(y)log2(p(y))$….(1) where p(y) is the marginal probability density function for the random variable Y. If the observed values of Y in the training data set S are partitioned according to the values of a second feature X, and the entropy of Y with respect to the partitions induced by X is less than the entropy of Y prior to partitioning, then there is a relationship between features Y and X. The entropy of Y after observing X is then:= $H(Y|X)=-\sum_{x\in X} p(x)\sum_{y\in y} p(y|x)log\ 2(p(y|x))$ ….(2)where p(y|x) is the conditional probability of y given x.

2.3.2. Information Gain

Given the entropy is a criterion of impurity in a training set S, we can define a measure reflecting additional information about Y provided by X that represents the amount by which the entropy of Y decreases. This measure is known as IG. It is given by IG = H(Y) – H (Y |X) = H (X) – H (X |Y)… (3) IG is a symmetrical measure (refer to equation (3)). The information gained about Y after observing X is equal to the information gained about X after observing Y. A weakness of the IG criterion is that it is biased in favor of features with more values even when they are not more informative.

2.3.3. Gain Ratio

The Gain Ratio is the non-symmetrical measure that is introduced to compensate for the bias of the IG. GR is given by $GR=\frac{IG}{H(X)}$…(4) As equation (4) presents, when the variable Y has to be predicted, we normalize the IG by dividing by the entropy of X, and vice versa. Due to this normalization, the GR values always fall in the range [0, 1]. A value of GR = 1 indicates that the knowledge of X completely predicts Y, and GR = 0 means that there is no relation between Y and X. In opposition to IG, the GR favors variables with fewer values.

2.4 Brief statistical analysis

Table 2
Distribution of data sets

Dataset Name	0	1	2	3	4	Total
Cleveland	164	55	36	35	13	303

Table 3
Ranking of attributes of attributes with information gain ratio.

S.No	Attributes:	Ranked
1.	13 tha	0.2249
2.	3 cp	0.2181
3.	12 ca	0.1949
4.	10 oldpeak	0.1569
5.	9 exang	0.1501
6.	8 thalach	0.1482
7.	11 slope	0.136
8.	1 Age	0.0722
9.	2 sex	0.058
10.	5 chol	0
11.	6 fbs	0
12.	4 trestbps	0
13.	7 restecg	0

3. METHODS DESCRIPTION

Here we select standard set of methods [13] for predicting from the data set described above. We consider three types of classifiers for our study such as tree based, Bayes approach based, and Meta level based classifiers. The following sections describe briefly the methods for classifying and results of such methods are tabulated further. Then final results are interpreted

3.1 Tree Classifiers

Supervised Learning is performed conducted using tree classifiers .We select four types of tree classifiers as shown below.

3.1.1 Decision Stump

A decision stump is a machine learning model consisting of a one-level decision tree. That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes. A decision stump makes a prediction based on the value of just a single input feature

3.1.2 J48

The first number is the total number of instances (weight of instances) reaching the leaf. The second number is the number (weight) of those instances that are misclassified. If your data has missing attribute values then you will end up with fractional instances at the leafs. When splitting on an attribute where some of the training instances have missing values, J48 will divide a training instance with a missing value for the split attribute up into fractional parts proportional to the frequencies of the observed non-missing values. This is discussed in the Witten & Frank Data Mining book as well as Ross Quinlan's original publications on C4.5.

3.1.3 Random Forest

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree

classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost, but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

3.1.4 REP Tree

Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back fitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces

3.2 Bayes Classifiers

These types of classifiers include probability measure for the class values and comes under supervised learning.

3.2.1 Bayes Net

Bayes Network learning using various search algorithms and quality measures. Base class for a Bayes Network classifier. Provides data structures and facilities common to Bayes Network learning algorithms like K2 and B.

3.2.2 Naïve Bayes

Class for a Naive Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data. For this reason, the classifier is not an Updateable Classifier you need the Updateable Classifier functionality, use the Naïve Bayes Updateable classifier. The Naïve Bayes Updateable classifier will use a default precision of 0.1 for numeric attributes when build Classifier is called with zero training instances.

3.2.3 Naive BayesUpdateable

Class for a Naive Bayes classifier using estimator classes. This is the updateable version of Naïve Bayes. This classifier will use a default precision of 0.1 for numeric attributes when build Classifier is called with zero training instances.

3.2.3 NaïveBayesMultinominal

Class for building and using a multinomial Naive Bayes classifier. A Bayesian network, Bayes network, belief network, Bayesian model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

3.3 Meta Classifiers

Most of the time, the aggregation of more than one classifier has better performance. Such combinational methods are shown below.

3.3.1 Adaboost

Class for boosting a nominal class classifier using the Adaboost M1 method. Only nominal class problems can be tackled. Often dramatically improves performance, but sometimes over fits.

3.3.2 Bagging

Class for bagging a classifier to reduce variance. Can do classification and regression depending on the base learner. Generate B bootstrap samples of the training data: random sampling with replacement. Train a classifier or a regression function using each bootstrap sample for classification: majority vote on the classification results. For regression: average on the predicted values. Reduces variation. Improves performance for unstable classifiers which vary significantly with small changes in the data set, e.g., CART. Found to improve CART a lot, but not the nearest neighbor classifier.

3.3.3 Stacking
Combines several classifiers using the stacking method. Can do classification or regression.

3.3.4 Logit Boost
This classifier is for performing additive logistic regression. This class performs classification using a regression scheme as the base learner, and can handle multi-class problems. Boosting is a machine learning meta-algorithm for reducing bias in supervised learning.

3.4 Function Classifiers
Most of the time, the aggregation of more than one classifier has better performance. Such combinational methods are shown below.

3.4.1 Logistic
Class for building and using a multinomial logistic regression model with a ridge estimator.

3.4.2 Multilayer Perceptron
A Classifier that uses back propagation to classify instances. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time. The nodes in this network are all sigmoid

3.4.3 RBF Network
Class that implements a normalized Gaussian radial basis function network. It uses the k-means clustering algorithm to provide the basis functions and learns either a logistic regression (discrete class problems) or linear regression (numeric class problems) on top of that. Symmetric multivariate Gaussians are fit to the data from each cluster. If the class is nominal it uses the given number of clusters per class. It standardizes all numeric attributes to zero mean and unit variance.

3.4.4 Simple logistic
Classifier for building linear logistic regression models. Logit Boost with simple regression functions as base learners is used for fitting the logistic models. The optimal number of Logit Boost iterations to perform is cross-validated, which leads to automatic attribute selection.

Table 4
Results with different classifiers for Cleveland dataset

Information Gain	Tree Classifiers	Bayes Classifiers	Meta Classifiers	Function Classifiers
0.058	57.4257	57.0957	58.0858	59.0759
0.0722	55.7756	60.726	58.7459	61.7162
0.136	55.7756	60.726	59.0759	61.3861
0.1482	56.4356	56.765	58.7459	61.0561
0.1501	56.7657	57.7558	57.7558	59.736
0.1569	56.9769	58.4158	60.726	58.4158

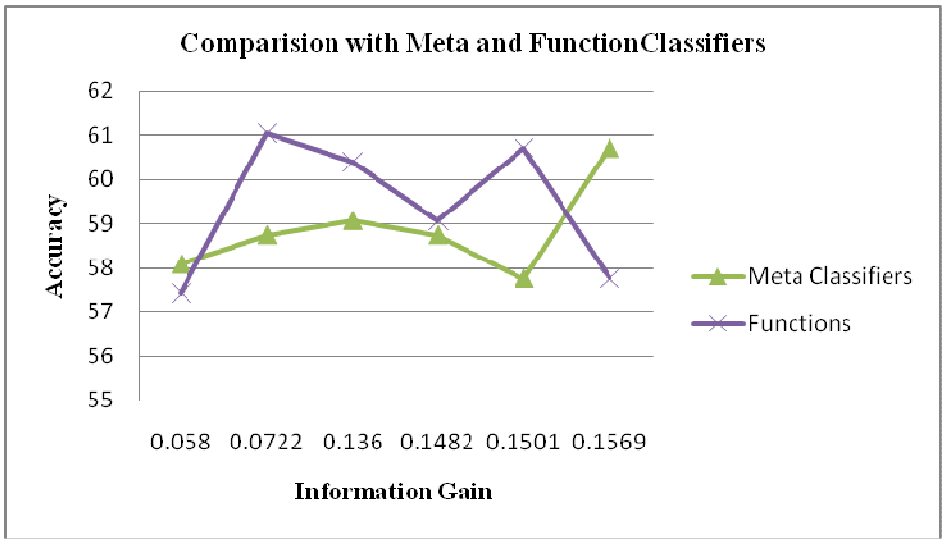


Figure 2
Threshold Information gain Vs Accuracy Part-I

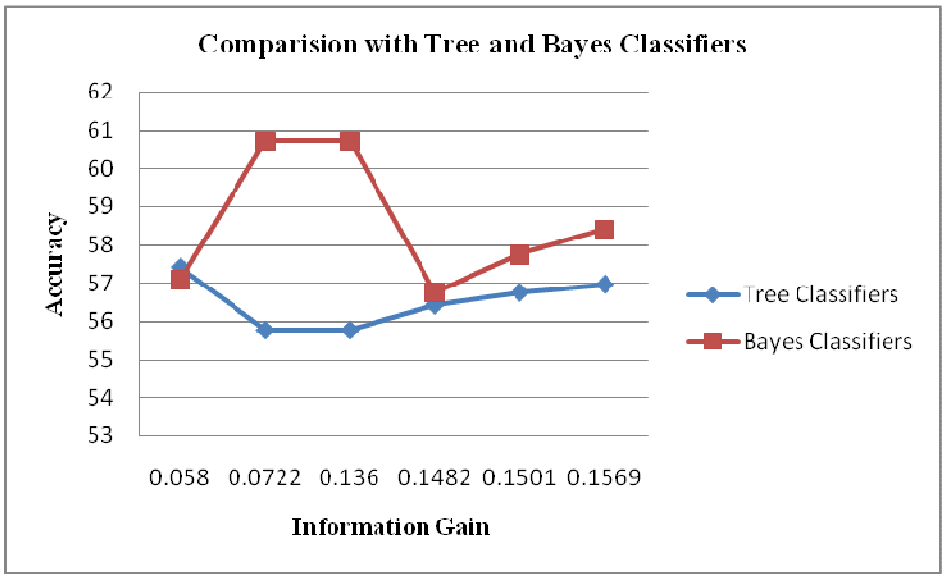


Figure 3
Threshold Information gain Vs Accuracy Part-II

4. RESULTS

Here we could increase the accuracy from 55%-61% by selecting attributes with their scores evaluated by information gain ratio. Figure 2 and 3 magnify the gap in the improved accuracies for the two parts namely Part I and Part II indicating {Meta, Function} classifiers, and{Tree, Bayes} classifiers respectively future we can extend this investigation for other types of datasets with different scoring methods. This paper the proof of concept for establishing the significance of reducing the search space by information gain ratio.

5. CONCLUSION

The effect of selection of attributes reflect in the accuracy of models for learning the heart disease data set is shown through the iteration over threshold values obtained by the

information gain. We could see the difference in the accuracy around 5%. However this can be verified with different data sets with more instances.

ACKNOWLEDGEMENT

The authors would like to thank the management of Bharath University for the support and encouragement for this research work.

REFERENCES

1. Companion slides for the text by Dr..H.Dunham, *Data Mining, Introductory andAdvanced Topics*, Prentice Hall, (2002).
2. Uci reprository https://archive.ics.uci.edu/ml/datasets.html downloaded on 13th Aug 2014.
3. A.Kumaravel, Pradeepa.R, Efficientmolecule reduction for drug design by intelligent search methods Int J Pharm Bio Sci; 4(2): (B) 1023 – 1029 Apr(2013)
4. Toward optimal feature selection using ranking methods and classification Algorithms Jasmina Novaković, Perica Strbac, Dusan Bulatović Yugoslav Journal of Operations Research, Number 1, 119-135 (2011)
5. https://www.waset.org/journals/waset/v68/v68-21.pdf world academy of science, engineering and technology, 2012.(Accessed on 9th july 2014)
6. A.Kumaravel and D.Udhayakumarapandian, Consruction Of Meta Classifiers For Apple Scab Infections Int J Pharm Bio Sci ; 4(4): (B) 1207 – 1213 Oct (2013).
7. A.Gelman, Y. S. Su, M.Yajima, J. Hill, M.Pittau, J. Kerman, and T. Zheng, "arm:Data Analysis Using Regression and Multilevel/Hierarchical Models," R package version 1.5-02.://CRAN.Rproject.org/package=arm,(2012).
8. A.Kumaravel and D.Udhayakumarapandian, A Novel Subset Selection For Classification Of Diabetes Dataset By Iterative Methods Int J Pharm Bio Sci 2014 July ; 5 (3) : (B) 1 - 8
9. L.Breiman,"RandomForests,"in*MachineLearning*, vol. 45, pp. 5-32, (2001).
10. http://ipm.ncsu.edu/apple/chptr5.html(Accessed on 3rd july 2014)
11. Dietterich, T. G., Jain, A., Lathrop, R.,Lozano-Perez, T. A comparison of dynamic reposing and tangent distance for drug activity prediction.Advances in Neural Information Processing Systems, 6. SanMateo, CA: Morgan Kaufmann.(B): 216--223(1994).
12. A.Stensvand, T. Amundsen, L. Semb, D.M.Gadoury, and R.C. Seem. Ascospore release and infection of apple leaves by conidia and ascospores of Venturia inaequalis at low temperatures.Phytopathology 87:1046-1053(1997).
13. Biggs, A. R. "Apple Scab" in Jones, A. L.,and Aldwinckle, H. S. (editors). Compendium of apple and pear diseases.APS Press, St. Paul, MN(1990).
14. http://weka.sourceforge.net/doc/weka/ (Downloaded on 10th Aug 2014).
15. Sourceboutwekahttp://www.cs.waikato.ac.nz/ml/weka/downloaded on 10 Aug 14
16. Performance comparison for geographically distinct datasets for heart disease Int J Pharm Bio Sci (To be published).
17. Comparative study of attribute selection using gain ratio and correlation based feature selection Asha Gowda Karegowda1, A. S. Manjunath2 & M.A.Jayaram3 International Journal of Information Technology and Knowledge Management Volume 2, No. 2, pp. 271-277 December (2010).
18. Toward optimal feature selection using ranking methods and classification Algorithms Jasmina Novaković, Perica Strbac, Dusan Bulatović Yugoslav Journal of Operations Research Number 1: 119-135 (2011).