Comparative Analysis of Feature Selection Algorithms for Medical Diagnosis

N.S.Nithya, Dr.K.Duraiswamy

Abstract— **Health care data contain large volume of valuable information for diagnosing diseases. The major challenge of the health care system is to extract useful patterns from these mass medical diagnosis data. In data mining technique, the feature selection method is used to extract relevant features in the original data set which contains noisy and irrelevant data. During the selection process, a decision criterion is used to remove irrelevant or redundant features. Elimination of irrelevant feature increases the prediction accuracy of classifying medical data and also decreases the computational time. To increase the performance and improve the design of classification algorithms, analyze the strength and weakness of the feature selection approaches are very essential. In this paper, the performance of the three filter based feature selection methods information gain, gain ratio and correlation have been analyzed. The selective feature selection methods are compared based on the classification accuracy for a given data set.**

Index Terms— **Data mining, Feature selection algorithms, Information Gain, Gain ratio, Correlation**

I. INTRODUCTION

Data Mining Techniques are used in many applications like e-business, Marketing, Fraud Detection and Health care management. Medical data mining has great potential for extracting the hidden patterns in the mass data sets of the medical field. Data mining technology provides a user friendly approach to the new and unknown patterns in the data. The medical expert system needs independent decision making in medical and engineering applications is growing, as data becomes easily available. The supervised learning algorithm contains the set of training instances called features and class label used to classify the diseases. The main objective of supervised learning is to maximize classification accuracy for an unseen data by mining relevant feature. The extracting essential feature is very important for medical data mining. Feature selection algorithms are used to extract those essential features in medical applications.

Feature selection methods are classified into the filter based approach, wrapper based approach and hybrid approaches. The filter based model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm [1, 2, 3]. The wrapper-based feature selection method [4], was used to identify the subset of relevant features combine with a classification algorithm that maximizes the performance of the prediction model. The feature subset selection algorithm comprises of the classification algorithm contain the evaluation criteria to find the best subset. The wrapper model needs one predefined mining algorithm and utilizes its performance as the evaluation criterion. It finds for features better suited to the mining algorithm aiming to enhance mining performance, but it also leads to be more computationally expensive than the filter model [5, 6]. The hybrid model combines the advantage of the two models by developing their different evaluation criteria.

In this paper some selective feature selection algorithm is applied to medical data sets to test its performance by accuracy comparison. The rest of the work is organized as Section 2 describes three feature selection algorithms and Section 3 experimental evaluation as well as comparative analysis of these algorithms. Conclusion and Future work is explained in Section 4.

II. FEATURE SELECTION ALGORITHMS

In machine learning feature selection is an important step for predictive modeling. The objectives of the feature selection process is to find the set of relevant features related to target and identifying the minimal subset of features that optimize prediction accuracy. It is important to know the difference between these two objectives in practice. For instance in biomedical fields the first objective is to find the potential causes of the diseases and the second objective is to eliminate the noise and improve the prediction accuracy. This is common in diagnosis system implementation and experiment design. Feature selection has been widely used to improve prediction accuracy of classifiers. The potential marker (causes of diseases) of diseases is more and more important for biomedical research and diseases study. The selection of optimal features is more complex compared to find optimal parameters for full set of features in the predictive modeling , first optimal feature subset is to be found and the model parameters are to be optimized [10]. Further section describes about selective filter based feature selection algorithms

A. Information Gain Measure

Information gain measure [7] is used for attribute selection. Based on the information gain assign the weight to different attribute and can get more accuracy in predictive modeling system like medical field etc. In any prediction model all attributes do not have same importance in predicting the class label. So different weights can be assigned to different attributes according to their information gain measure. Attribute with highest information gain [8] chosen as the test

Manuscript received Dec 09, 2014

N.S.Nithya, Department of Computer Science and Engg, K.S.R.College of Engineering, India
Dr.K.Duraiswamy, Department of Computer Science and Engg K.S.Rangasamy College of Technology,India

attribute and assign highest weight value. Information theoretic approach minimizes the number of tests needed to classify an object.

The entropy H(D) of a dataset D is a measure of the disorder/variation/information in it. If all the records in the dataset belong to the same class, then the entropy would be zero. If all the records are uniformly distributed among the different classes, the entropy would be maximized. The entropy H(D) of a dataset D whose records are divided into m classes with probabilities $p_1,p_2…..p_m$ is defined as

$$H(D) = - \sum_{i=1........m} p_i \log p_i \qquad (1)$$

The best split for each attribute is chosen based on a criterion known as information gain. Given that a dataset D is split into $D_1,D_2,D_3……Dn$ the information gain of the split is computed as

$$\text{Gain}= H(D) \text{ - } \sum_{i=1}^{n} P(D_i)\, H(D_i) \qquad (2)$$

In this equation, the first part is the entropy of the dataset before the split, whereas the second part is the average (or) expected entropy of the collection of the datasets after the split. The ID_3 algorithm selects the split with maximum information gain. The medical dataset mostly contain categorical attributes. To make this approach feasible, the ID_3 algorithm only considers categorical attributes because the number of distinct values is small and hence, can be enumerated.

B. Gain Ratio Based Feature Selector

Split method is most important component of decision tree learner. To attain high predictive accuracy for many situations, split method (information gain ratio) is the best one. The information gain measure is biased towards tests with many outcomes. The major drawback of using information gain is that it tends to choose attributes with large numbers of distinct values over attributes with fewer values even though the later is more informative [9]. For example consider an attribute that is name of the disease in patient database. A split on disease name would result in a large number of partitions; as each record in the database has a different name for different patient. So the information required to classify database with this partitioning would be nearly a small value clearly, such a partition is useless for classification.

C4.5, a successor of ID3 [10], uses an extension to information gain known as gain ratio (GR), which attempts to overcome the bias. The WEKA [11] classifier package has its own version of C4.5 known as J4.8. We use J4.8 to identify the significant attributes. Let D be a set consisting of d data samples with n distinct classes. The expected information needed to classify a given sample is given by

$$I(D) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (3)$$

where p_i is the probability that an arbitrary sample belongs to class Ci. Let attribute A have v distinct values. Let dij be number of samples of class Ci in a subset Dj. Dj contains those samples in D that have value aj of A. The entropy based on partitioning into subsets by A, is given by

$$E(A) = -\sum_{i=1}^{n} I(D) \frac{d_{1i}+d_{2i}+\cdots+d_{mi}}{d} \qquad (4)$$

The encoding information that would be gained by branching on A is

$$Gain(A) = I(D) - E(A) \qquad (5)$$

C4.5 applies a kind of normalization to information gain using a “split information” value defined analogously with Info (D) as

$$Splitinfo_A(D) = -\sum_{j=1}^{v} \left(\frac{|D_j|}{|D|}\right) \log_2 \left(\frac{|D_j|}{|D|}\right) \qquad (6)$$

This value represents the information computed by splitting the dataset D, into v partitions, corresponding to the v outcomes of a test on attribute A [12]. For each possible outcome, it considers the number of tuples having that outcome with respect to the total number of tuples in D. The gain ratio is defined as

$$GainRatio(A) = \frac{Gain(A)}{Splitinfo(A)} \qquad (7)$$

The attribute with maximum gain ratio is selected as the splitting attribute. WEKA data mining tool [15] provides the environment to calculate the information gain ratio.

C. Correlation based feature selector

Correlation based feature selector rank the feature based on correlation based heuristic evaluation function. The evaluating criteria of subset features based on features highly correlated with the class and uncorrelated with each other. For example, consider prediction of diseases depends on the potential causes of the diseases which are highly correlated with that classification of diseases. The potential attribute is considered as a relevant feature. Irrelevant features should be eliminated since they have low correlation with the class and redundant features also eliminated because they highly correlated with remaining features. The selection of a feature will depend on how it predicts classes in areas of the instance space not already predicted by other features. Correlation based feature evaluation function is given as follows:

$$r_{cc} = \frac{k\overline{r_{ce}}}{\sqrt{k+k(k-1)\overline{r_{ee}}}}; \qquad (8)$$

Where r_{cc} is the rank correlation between feature subset and the class variable containing k features, $\overline{r_{ce}}$ is the mean feature- class correlation and $\overline{r_{ee}}$ is the feature-feature inter-correlation. The numerator of the equation indicates prediction of feature and denominator indicating the redundancy among features. From this measurement we obtain a set of ranked features.

III. DATASETS AND EXPERIMENTAL RESULTS

To verify the accuracy of the above selective three feature selection algorithm experiments have been performed on the datasets collected from the UCI repository. The data set obtained from the UCI repository are 10 input features (attributes), 2 classes and 699 samples of Wisconsin breast cancer data, 13 input features (attributes), 2 classes and 303 samples of Heart disease data, 4 input features (attributes), 2 classes and 150 samples of Iris data, 10 input features (attributes), 2 classes and 345 samples of Liver and 8 input features (attributes), 2 classes and 768 samples of Pima Indian diabetes datasets are used to test the effectiveness of feature selection algorithms. The following table-1 shows the feature selected by three filter based selective algorithm. The results have been obtained from the Weka Tool. The input files to the WEKA are datasets that is used here in CSV format.

Table-1 Features Selected by Filter Based Algorithms

Dataset	**Feature set**	**Information Gain**	**Gain Ratio**	**Correlation**	**Correlation and GainRatio**
Heart disease	13	7	7	7	7
Liver	10	5	5	6	5
Breast cancer	9	5	6	8	7
Pima Indian Diabetes	8	4	4	4	4
Iris	4	2	2	2	2

The performances of the three selective algorithms are evaluated by prediction accuracy. Most relevant and similar features are selected by both correlation and gain ratio in five data set .The information gain select the potential feature different from the above two but the number of feature selection is more or less same for three algorithms. So the combined approach of correlation and gain ratio based feature selection is also evaluated which gives an optimal feature selection as shown in Fig-1. Accuracy of the correlation and gain ratio based approach gives better than the three selective algorithms as shown in Fig-2.

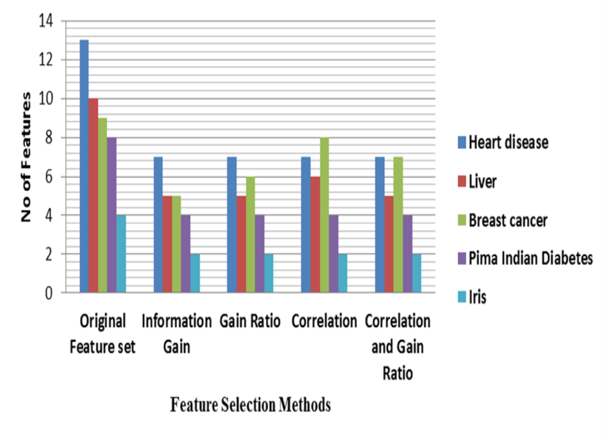

Fig-1 Comparative analysis of feature selected by three feature selection algorithm

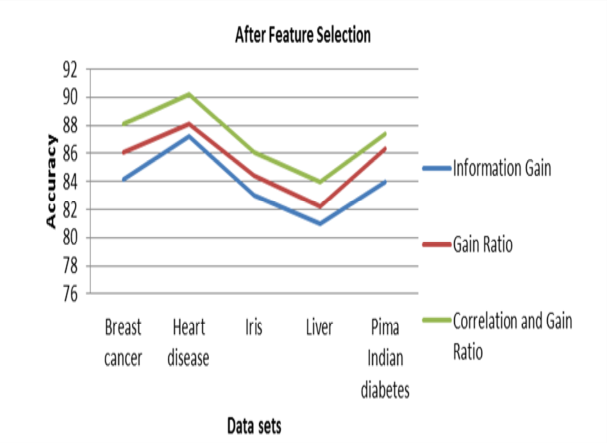

Fig-2 Comparative analysis of accuracy for five data set.

CONCLUSION

In this paper selective filter based feature selection algorithms were considered and the performance evaluated in terms of prediction accuracy. The three selective filter based feature selection algorithms such as information gain, gain ratio and correlation methods have been analyzed with the five medical dataset. The three selective algorithms select the same potential attributes for Pima Indian Diabetes and Iris data set .The correlation and gain ratio methods select the same relevant features for predicting diseases more or less similar for heart disease, Breast cancer and liver data set also. So the hybrid approach using correlation and gain ratio has been applied to feature selection process which gives an optimal solution. Accuracy of correlation and gain ratio based feature selection is better compared to information gain and gain ratio feature selection. With the above data set correlation and gain ratio based approach is giving better results with the other selective filter based feature selection algorithms.

REFERENCES

[1] [M. Dash, K. Choi, P. Scheuermann, and H. Liu. “Feature selection for clustering – a Filter solution”. In Proceedings of the Second International Conference on Data Mining, pages 115–122, 2002.

[2] M.A. Hall. “Correlation-based feature selection for discrete and numeric class machine learning”. In Proceedings of the Seventeenth International Conference on Machine Learning, pages 359–366, 2000.

[3] L. Yu and H. Liu. Feature selection for high dimensional data: a fast correlation-based filter solution. In Proceedings of the twentieth International Conference on Machine Learning, pages 856–863, 2003.

[4] Hsueh-Wei Chang, Yu-Hsien Chiu, Hao-Yun Kao, Cheng-Hong Yang, and Wen- Hsien Ho “Comparison of Classification Algorithms with Wrapper-Based Feature Selection for Predicting Osteoporosis Outcome Based on Genetic Factors in a Taiwanese Women population” International Journal of Endocrinology, Volume 2013, Article ID 850735, 8 pages

[5] J. G. Dy and C. E. Brodley. Feature subset selection and order identification for unsupervised learning. In Proceedings of the Seventeenth International Conference on Machine Learning, 2000 pages 247–254.

[6] Y. Kim, W. Street, and F. Menczer. Feature selection for unsupervised learning via evolutionary search. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000, pages 365–369.

[7] Vikram Pudi, P.Radha Krishna, Data Mining, Oxford University Press 2009 Edition

[8] M. Ashraf, Kim Le, Xu Huang, Information Gain and Adaptive Neuro-Fuzzy Inference System for Breast Cancer Diagnoses, pp.911-915

[9] Asha G. K, A. S. Manjunath, M. A. Jayaram, “A Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection”, International Journal of Information Technology and Knowledge Management, July – December 2012, Volume 2,pages 271 – 277.http://www.cs.waikato.ac.nz/~ml/weka/.

[10] J. R. Quinlan, “Induction of decision tress”, Machine Learning, Kluwer Academic Publishers, 1986, pages 81-106.

[11] J. Han and M. Kamber, Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffm ANN Publishers (2001).