₹ 200 ISSN - 2249-555X
Volume : 1 Issue : 9 June 2012

Journal for All Subjects

www.ijar.in

Listed in International ISSN Directory, Paris.

ISSN - 2249-555X

Indian Journal of Applied Research

Journal for All Subjects

Editor,
Indian Journal Of Applied Research
8-A, Banans, Opp. SLU Girls College, New Congres Bhavan, Paldi,
Ahmedabad-380006, Gujarat, INDIA
Contact.: +91-9824097643 E-mail : editor@ijar.in

INDEX

Volume : 1 | Issue : 9 | June 2012 ISSN - 2249-555X

Research Paper Computer Science

Hybrid Attribute Selection Process for Decision Tree Based Classification Algorithms

* Mr. A. Jebamalai Robinson ** Mrs. S. C. Punitha
*** Dr. P. Ranjit Jeba Thangaiah

* Lecturer, Department of Comp. Applications, Karunya University, Coimbatore

** Head of the Department, PSGR Krishnammal College for Women, Peelamedu

*** Assistant Professor (SG), Department of Comp. Applications, Karunya University, Coimbatore

ABSTRACT

Feature selection can be defined as techniques available for reducing inputs to a manageable size for easy and fast processing and analysis. It is a mandatory step in all data mining tasks involving large datasets. In the present study, a novel method of selecting significant features is proposed which combines the result of three frequently used traditional feature selection algorithms. The three algorithms are correlation-based method, consistency subset evaluation method and Wrapper Subset evaluation method. The result of feature selection is analyzed using two classifiers, namely, J48 and Random Forest. Experimental results with respect to number of features selected and classification accuracy proved that the proposed method is efficient in selecting features that are most important for classification.

1. INTRODUCTION

The beginning of the twenty first century has brought considerable advances in the field of computer-based information retrieval systems, where data with "hidden asset" called "knowledge" is quickly becoming the most valuable resource. The rapid growth in the number and size of databases, dimension and complexity of data has made it necessary to automate the analysis process, whose results can then be used by decision-making processes. According to Fayyad and Uthurusamy (2002), the ability to capture data has outpaced the ability to process and acquire meaningful knowledge from it. Moreover, the amount of data available far exceeds the analyzing power of the humans. Thus, the demand for knowledge discovery techniques and mining techniques grows tremendously and is considered to be very important in current information explosion environment. **Data Mining uses different varieties of data analysis tools to discover knowledgeable patterns and relationships in data**.

Different kinds of patterns are discovered accordingly based on the data mining tasks employed. Examples include clustering, classification, regression and summarization. One important step in all these tasks is the Feature Selection (FS) process. The number of records and attributes present in modern databases are very large and a straight forward manner of manipulating these data has disadvantages like huge computation and large storage space, which indirectly increases the cost during knowledge recovery. Thus, a FF procedure is always desired. It is defined as the tools and techniques available for reducing inputs to a manageable size for easy and fast processing and analysis.

FS has been an active research area in data mining, where the goal is to find a subset that increases the classification accuracy. The main aim of this paper is to propose a fusion model for efficient FS and analyze its performance on classification. Three FS processes, namely, correlation-based method, consistency subset evaluation method and Wrapper Subset evaluation method are combined to build a fusion model, from which the correlated attributes that have impact on classification are selected. The applicability of using Random Forest (RF) and J48/C4.5 classifiers on the fused feature set is analyzed.

2. REVIEW OF LITERATURE

Antipov and Pokryshevskaya (2011) proposed a systematic mass appraisal approach for residential estate and detected under or overestimated parameters using RF algorithm for residential apartments. The best split procedure was enhanced to use the random sample of m variables rather than using the predictors. The system was compared with many traditional algorithms and proved to produce accurate results. Xie et al. (2009) proposed a novel learning method called Improved Balanced RF (IBRF). To overcome the problem of imbalance in standard RF, two methods namely, cost-sensitive weighted RF and balanced RF were proposed and combined to form a third hybrid method. Results proved that the hybrid model produced better results.

Harb et al. (2009) explored the usage of Random Forest in order to categorize and rank the characteristics like drivers/ vehicles/ environments for a crash avoidance maneuvers. Random Forest was also used as a source for unveil the attributes of drivers, vehicles, and environments which is associated with crash evasive actions, 'no avoidance maneuver' and 'with avoidance maneuver', and also ranked the importance of the characteristics on accident evasive maneuvers while eliminating potential effects of correlated inputs.

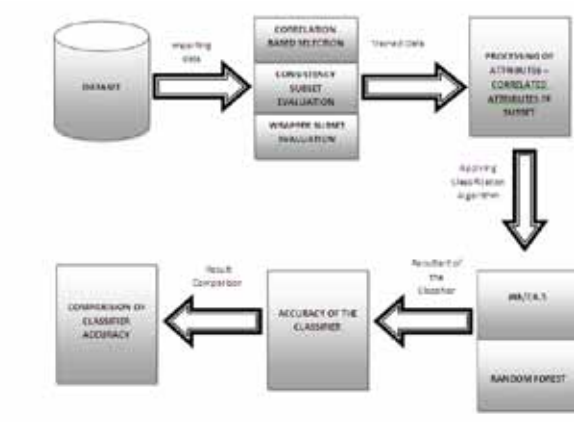
Smith et al. (2010) performed a Bacterial Source Tracking (BST) analysis using two statistical techniques, Random Forest and discriminant analysis (DA) using antibiotic resistance profiles. The aim was to identify the sources for the fecal contamination of a Texas water body. This study is the first known application which was published about Random Forest in the field of BST. Verikas et al (2011) presented a framework which was generally meant for designing a fuzzy rule-based classifier. Two-Stage genetic search method is used to evolve the structure and parameters of the classifier. SOM tree algorithm is used to reduce the search space, by constrained the classifier structure as a tree structure. In the genetic search process the salient input variables which are specific for each fuzzy rule and are found. Computer simulations showed about the eliminating large number of rules and input variables from the model without deteriorating the classification accuracy. Stumpf and Kerle (2011) used RF for classifying landslide mapping from variety of optical sensor to identify landslide and non-landslide objects. Random clas-

sifier trained with 20% of all land and non-landslide objects, yielded higher accuracy in the classification.

Gunes and Polat (2009) combined C4.5 decision tree classifier and one against-all method to overcome multiclass problems in terms which also helps as a solution for classification accuracy.The results showed that the proposed method has produced very promising results in the classification of multi-class problems. Song et al. (2008) presented a system for predicting missing data values in six real world software project database using k-NN missing data imputation and C4.5 classifier, the results tells clearly that prediction accuracy can be improved with the help of k-NN imputation.

3. METHODOLOGY

The methodology used in the proposed algorithm is given in Figure 2.



The proposed algorithm was developed in Weka environment and its involves the following steps. After importing the dataset, the first step uses three FS algorithms namely Correlation Based Feature Selection, Consistency Subset Evaluation and Information Gain Attribute Evaluation to select only those attributes that have relevance to classification. Information gain metric is used to select highly relevant features from the datasets. As the features selected are used for classification, care was taken to consider the target attributes and classification algorithm type. The second step then applies these features for classification. Two classifiers, namely, RF and C4.5, are used for this purpose. The description on the selected FS algorithms is given below (www.opentox.org/dev/documentation/components/cfsfeatsel).

The Correlation Based Feature Selection (CBFS) technique is a correlation-based filter method that gives high scores to subsets that include futures that are highly correlated to the class attribute but have low correlation to each other. Let S be an attribute subset that has k attributes [opentox.org] . InfoGain Attribute Evaluation (IAE) evaluates the worth of an attribute by measuring the information gain with respect to the class. InfoGain(Class,Attribute) = H(Class) – H(Class | Attribute), where H is the information entropy [opentox.org]. Consistency Subset Evaluation (CSE) method determines the worth of a subset of attributes by the level of consistency in the class values when the training instances are projected onto the subset of attributes. Consistency of any subset can never be lower than that of the full set of attributes, hence the usual practice is to use this subset evaluator in conjunction with a random or exhaustive search which looks for the smallest subset with consistency equal to that of the full set of attributes.

The proposed system then builds a Feature Selection Fusion (FSF) algorithm using steps given in Figure 2. The fusion algorithm is a procedure which uses a fusion method to select important features from the result of each of the selected FS algorithm. The resultant step will produce a feature set which has key features that will have great impact on classification.

The fusion algorithm is a proposed algorithm which consists of two step process, where the first step selects the attributes which are common among the resultant attributes from the above three algorithms and the second step selects the left out attributes as uncommon attributes. The relevant features from these are selected using a threshold. Pseudo code of FSF is given below

Algorithm Attribute Selection Fusion
Input: Attribute subset from 3 Methods
Scan the attribute subset from the first method
Scan the attribute subset from the second method
Check for the commonality of the attributes from first two methods and form an attribute subset
Scan the attribute subset from the third method
Check for the commonality of the attributes from the first two methods as well as with the resultant of the selected attributes of the subset and form a correlated attribute subset
Check for the attributes that are ranked and not part of the first two methods
Keep a threshold condition and take those attributes and attach it to the correlated subset of attributes
Output: Correlated Subset of Attributes

Fig 2: Pseudo code of FSF

The last step takes the result of attribute fusion algorithm as input to the selected two classifiers and groups them according to predefined categories.

4. RESULTS AND DISCUSSION

The proposed attribute selection feature fusion algorithm was developed using prototype in Weka 3.6. Three datasets are taken from the machine learning UCI Machine Learning Repository (http:// WEKA Datasetselected/AutoUniv/AutoUniv.htm). First dataset is auto_univ with 10,000 instances and 251 attributes, second is adult dataset with 32,561 instances and 14 attributes and third is arrhythmia with 452 instances and 280 attributes. Two performance metrics, namely, number of attributes selection and classification accuracy were used. Table 1 presents the number of attributes selected by the selected attribute selection algorithm along with proposed hybrid method.

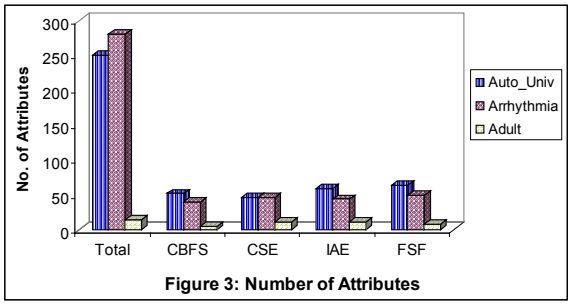


Figure 3: Number of Attributes

From the table, it can be seen that with all the three datasets, the best performance was produced by the proposed FSF algorithm. The selection of attributes by FSF is small when compared among the three other algorithms.

Table --- shows the classification accuracy of the traditional and proposed feature selected algorithms while using J48 and RF classifiers.

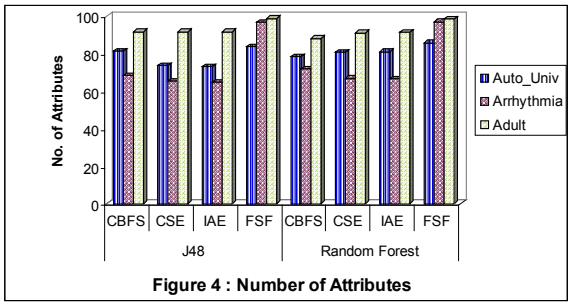


Figure 4 : Number of Attributes

The maximum accuracy achieved through FSF algorithm with both J$8 and RF classifier proves the efficiency of the fusion technique used while selecting the significant featues from a dataset. All the three selected datasets showed similar trend. The various results show that the method that performs fusion selection on the selected attributes produces better result and hence is a better candidate for classification problems.

5. CONCLUSION

This paper presented a fusion algorithm for attribute selection that can enhance the classification accuracy while reducing the computation time required. The work focused on three techniques, namely, Correlation Based Future Selection, Consistency Subset Evaluation and Information Gain Attribute Evaluation and two classifiers, J48 and RF. Performance evaluation using number of attributes selected and accuracy showed that the proposed model produces improved results when compared with the traditional algorithms. In future, more FS method will be combined with the proposed algorithm and the aspect of implementing parallel processing for the various steps are also considered.

REFERENCES

Antipov, E.A. & Pokryshevskaya, E.B. (2011) Mass appraisal of residential apartments: An application of Randomforest for valuation and a CART-based approach for model diagnostics, Expert Systems with Applications, 39(2), 1772–1778 | Fayyad, U. and Uthurusamy, R. (2002) Evolving data into mining solutions for insights, Communications of ACM, Vol. 45, Issue 8, Pp. 28-31. | Gunes, S. & Polat, K. (2009) A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems, Expert Syst. Appl., 36(2), 1587-1592. | Harb, R., Yan, X., Radwan, E. & Su, X. (2009) Exploring precrash maneuvers using classification trees and random forests, Accident Analysis & Prevention, 41(1), 98–107 | Smith, A., Sterba-Boatwright, B. & Mott, J. (2010) Novel application of a statistical technique, Random Forests, in a bacterial source tracking study, Water Research, 44(14), 4067–4076 | Song, Q., Sheppard, M., Chen, X. & Liu, J. (2008) Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation, Journal of Systems and Software, 81(12), 2361-2370. | Stumpf, A. & Kerle, N. (2011) Combining Random Forests and object-oriented analysis for landslide mapping from very high resolution imagery, Procedia Environmental Sciences, 3, 123–129 | Verikas, A., Guzaitis, J., Gelzinis, A. & Bacauskiene, M. (2011) A general framework for designing a fuzzy rule-based classifier, Knowledge and Information Systems, 29(1), 203-221. | Xie, Y., Li, X., Ngai, E.W.T. & Ying, W. (2009) Customer churn prediction using improved balanced random forests, Expert Syst. Appl., 36(3), 5445-5449. | Opentox.org | Weka.sourceforge.net

Sara Publishing Academy
Indian Journal Of Applied Research
Journal for All Subjects

Editor,
Indian Journal Of Applied Research
8-A, Banans, Opp. SLU Girls College,
New Congres Bhavan, Paldi, Ahmedabad-380006.
Contact.: +91-9824097643 E-mail : editor@ijar.in

Printed at Unique Offset, Novatsing Rupam Estate, Opp. Abhay Estate, Tavdipura, Shahibaug, Ahmedabad