Volume: 09 Issue: 06 Pages: 3640-3644 (2018) ISSN: 0975-0290

Classification Algorithms with Attribute Selection: an evaluation study using WEKA

Dr Gnanambal S
Department of Computer Science, Raja Dorai Singam Govt Arts College, Sivagangai
Email: gnanardm@gmail.com
Dr Thangaraj M
Department of Computer Science, Madurai kamaraj University, Madurai
Email: thangarajmku@yahoo.com
Dr Meenatchi V.T
Department of CA & IT, Thiagarajar College, Madurai
Email: vtmeenatchi@gmail.com
Dr Gayathri V
Department of CA, NIT, Tiruchi
Email: debuggay3@gmail.com

ABSTRACT

Attribute or feature selection plays an important role in the process of data mining. In general the dataset contains more number of attributes. But in the process of effective classification not all attributes are relevant. Attribute selection is a technique used to extract the ranking of attributes. Therefore, this paper presents a comparative evaluation study of classification algorithms before and after attribute selection using Waikato Environment for Knowledge Analysis (WEKA). The evaluation study concludes that the performance metrics of the classification algorithm, improves after performing attribute selection. This will reduce the work of processing irrelevant attributes.

Keywords – **attribute filters, attribute selection, classification, data mining, Weka**

Date of Submission: April 04, 2018 Date of Acceptance: April 18, 2018

I. INTRODUCTION

Classification is the one of the main technique used for discovering pattern from known classes [1]. In real word, dataset contains hundreds of attributes. But not all the attributes are needed to complete the mining task [2]. In order to find the importance of attributes, feature selections algorithms are utilized. Instead of processing all the attributes, only relevant attributes are involved in the mining process. This will reduce processing time as well as increase the performance of mining task. Therefore attribute selection algorithms are applied before applying data mining tasks such as classification, clustering, outlier analysis and so on.

Attribute selection is a two step process one is subset generation and another one is ranking. Subset generation is a searching process which is used to compare the candidate subset to the subset already determined [3]. If the new candidate subset returns better results in terms of certain evaluation then the new subset is termed as best one. This process is continued until termination condition is reached.

The next one is Ranking of attributes which is used to find the importance of attributes [4]. There are many ranking methods such as which are mostly based on statistics or information theory. There are two varieties of attributes selection algorithms. i) Filter approach ii) Wrapper approach. The learning algorithms itself uses the attribute selection task then it is called wrapper approach [5]. In filter approach the attributes are evaluated on the basis of evaluation metrics with respect to the characteristics of the dataset [6]. The Table 1 lists the comparison between filter and wrapper approach in terms of computational time, attribute dependencies and so on.

Table 1: Comparison between filter and wrapper approach

	Filter	***Wrapper***
Computational time	Simple and fast	Complex and Slower
In terms of attribute dependencies	Only to Some degree	Fully incorporated
Cost	Less expensive	Expensive
Scaling ability to high dimensional dataset	Easy	Complex

This research article is organized as five sections: Of which, Section 1 is the introduction to attribute selection and its approaches, Section 2 is devoted to the related literatures, Section 3 presents the adopted dataset and algorithms utilized. Section 4 deals with the Experimental results of classifier with respect to precision, recall and F-measure before and after attribute selection. Section 5 records the conclusion.

Volume: 09 Issue: 06 Pages: 3640-3644 (2018) ISSN: 0975-0290

II. RELATED LITERATURES

This paper [7] presents the introduction about the various classification and feature selection techniques frequently used data mining. It also states the importance of filter and wrapper approaches of feature selection methods. But this study does not contribute to any experimental study.

This study [8] is applied to predict studence performance. For that purpose feature selection techniques such as Chi-square, InfoGain, and GainRatio are utilized. Then classification task is carried out by the use of NBTree, MultilayerPerceptron, NaiveBayes and Instance based –K- nearest neighbor classifiers. The result concludes that the accuracy of the prediction was improved because of the applied filter techniques. But there is no comparative study presented to compare the filter and wrapper approaches.

This comparative study [9] determines the most relevant subset of attributes based minimum cardinality. In order to find the goodness of features, the six feature selection algorithms are involved in this study. It can be measured in terms of F-measure and ROC value. The result assures that the computational time and cost is decreased with minimum number of features.

There are many number feature selection algorithms available today. Hence this paper [10] presents the benefits and drawbacks of the some feature selection algorithms in terms of efficiency. Nearly 12 feature selection algorithms are involved in this study.

III. DATASET AND OUTCOME OF PREDICTION

The main goal of this paper is to apply the filter approach in a credit dataset of Germany. This dataset contains 21 attributes. The attributes of the germany_credit dataset is listed in Table 2.

This evaluation study is implemented in the data mining tool, WEKA . The three filter approaches are applied to the dataset i) CfsSubsetEval (CSE) ii) CorrelationBased Feature Selection (CB) iii) GainRatio Attribute evaluation (GR) iv) InformationGain Attribute evaluation (IG). These four approaches are explained below.

CfsSubsetEval (CSE):

This method measures the significance of attributes on the basis of predictive ability of attributes and its degree of redundancy. The subsets which are having less inter-correlation but highly correlated to the target class are preferred. This attribute evaluator with the Breadth First Search is applied to the german_credit dataset. The ranking of first five attributes are taken into accounts which are 1. checkingstatus, 2. duration, 3.history, 4. credit-amt, 5. savings-status. Fig 1 shows the ranking of attributes with respect to CfsSubsetEval method.

Correlation Attribute Eval (CA):

CA evaluates the attributes with respect to the target class. Pearson's correlation method is used to measure the correlation between the each attributes and target class attribute. It considers nominal attributes in value basis and each value acts as an indicator. . By the combination of this attribute evaluator with the Ranking method of Search is applied to the german_credit dataset. The ranking of first five attributes are taken into accounts which are 1.checkingstatus, 2.duration, 3.credit-amt, 4.savings-status and 5.housing. Fig 2 shows the ranking of attributes with respect to Correlation Attribute Eval method.

Table 2: Attributes list of german_credit dataset

No	Attribute name	Range
1	checking_status	{<0|>=200|no checking|<200}
2	Duration	4 to 60
3	credit_history	'critical'|'other existing credit'| 'existing paid'| 'delayed previously'| 'no credits'|'all paid'
4	Purpose	radio/tv| education| furniture/equipment|new car|old car| business| repairs
5	credit_amount	250 to 18424
6	savings_status	<100|>500|<1000| 'no known savings'|>1000
7	Foreign_worker	{yes|no}
8	Housing	{own|rent|for free}
9	employment	1 to 7
10	installment_commitment	1 to 4
11	personal_status	{'male single'|'female div/dep/mar'|'male div/sep'|'male mar/wid'}
12	other_parties	{none|guarantor}
13	residence_since	1 to 4
14	property_magnitude	{'real estate'|car|'life insurance'|'no known property'
15	Age	21 to 67
16	other_payment_plans	{bank|none}
17	existing_credits	1 or 2
18	job	{'unskilled resident'|'skilled'| 'high qualif/self emp/mgmt'|
19	num_dependents	1 or 2
20	own_telephone	Yes or no
21	class	Good|bad

Volume: 09 Issue: 06 Pages: 3640-3644 (2018) ISSN: 0975-0290

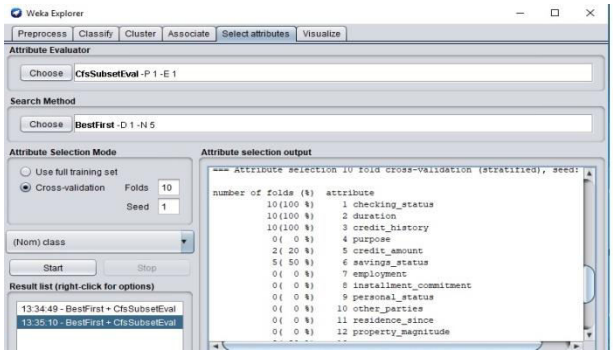

Fig 1: Ranking of attributes with respect to CfsSubsetEval method

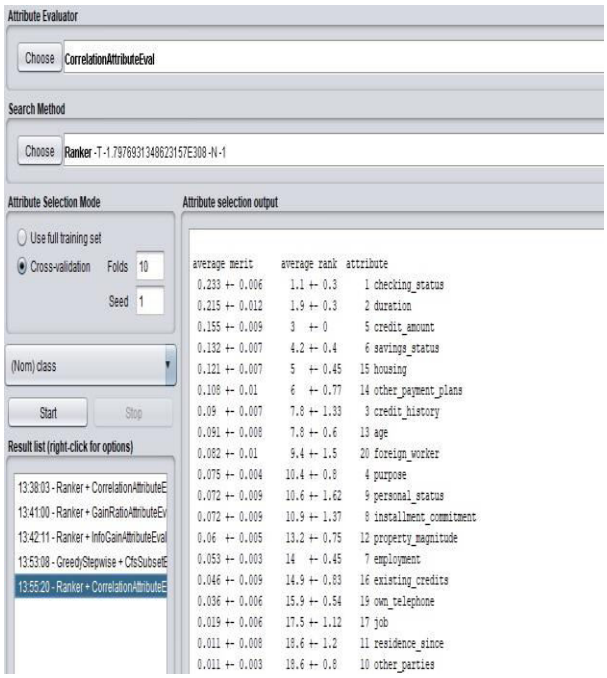

Fig 2: Ranking of attributes with respect to Correlation Attribute Eval (CA) method

GainRatio Attribute evaluation (GA):

This method measures the significance of attributes with respect to target class on the basis of gain ratio. It can be calculated by the following formula,

GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute)

Where H represents the Entropy. This attribute evaluator with the Ranker Searching is applied to the german_credit dataset. The ranking of first five attributes are taken into accounts which are 1. checkingstatus, 2.duration, 3.history, 4.forien and 5.credit-amt. Fig 3 shows the ranking of attributes with respect to GainRatio Attribute evaluation method.

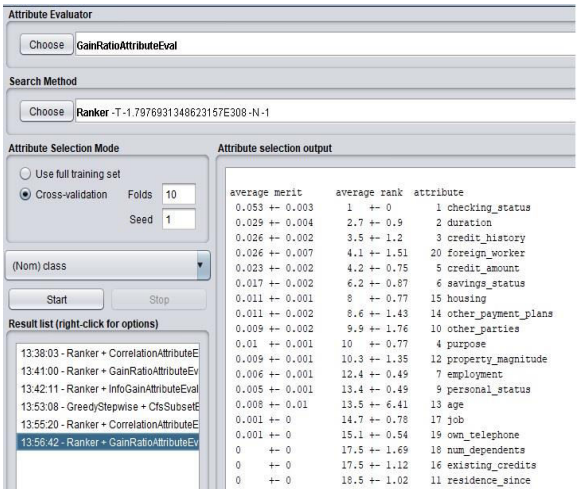

Fig 3: Ranking of attributes with respect to GainRatio Attribute evaluation method

Information Gain Attribute evaluation (IG):

This method measures the significance of attribute by the measure of information gain calculated with respect to target class. It can be calculated by the formula,

InfoGain(Class,Attribute) = H(Class) - H(Class | Attribute)

Where H represents the Entropy. By the combination of this attribute evaluator with the Ranking method of Search is applied to the german_credit dataset. The ranking of first five attributes are taken into accounts which are 1.checkingstatus, 2.duration, 3.credit-amt, 4.savings-status and 5.housing. Fig 4 shows the ranking of attributes with respect to InformationGain Attribute evaluation method.

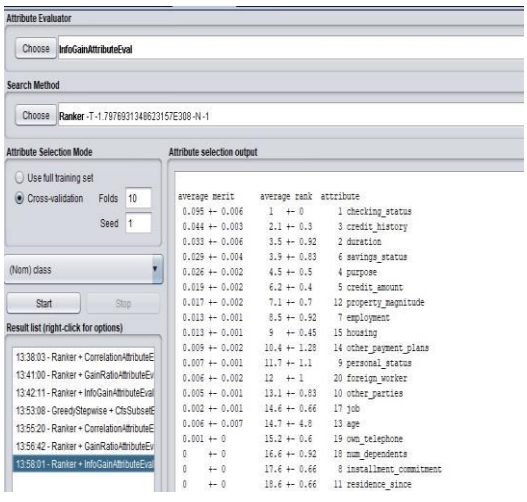

Fig 4: Ranking of attributes with respect to InformationGain Attribute evaluation method

This aforementioned attribute selection algorithms results the ranking of attributes. Based on the returned ranking of attributes, the dataset is modified. That is, the irrelevant attributes are removed. This modified dataset is involved into the classification task. Then the

Volume: 09 Issue: 06 Pages: 3640-3644 (2018) ISSN: 0975-0290

result of classifier before and after attribute selection is compared. For that purpose the PART classifier is utilized. The summary of the PART learning algorithm is given below.

PART is a decision list generation algorithm. It is a clone of C4.5 decision tree learner [11]. PART generates more number of rules than other algorithms. The main feature of PART is that it doesn't contain the global optimization phase. Instead PART uses separate and conquer strategy to create rules. The PART algorithm utilizes the pessimistic pruning. PART algorithm generates a decision tree and it consists of branches to unexpanded subtrees. In PART the tree construction and pruning operations are combined to produce the subtree which cannot be expanded further. A rule is derived from a partial tree. Each leaf drives a rule and the best leaf is selected from which leaf that travels more number of instances is expanded into leaves.

IV. EXPERIMENTAL RESULTS

The PART classifier is first applied to the german_credit dataset then the classifier is applied to modified dataset with respect to attribute evaluator. Then compare the results of PART with original dataset and the modified dataset after selecting the attributes with respect to the four attribute evaluator methods. The Fig 5 shows the results of PART classifier with respect to the original german_credit dataset. The Table 3 presents the results of PART classifier after applying the four attribute evaluators with respect to Precision, Recall, F-measure and Mean absolute error. The results shows that the combination of PART classifier with CfsSubsetEval attribute evaluator performs well in terms of precision, recall, f-measure. This CfsSubsetEval method also reduces the mean absolute error of the PART classifier.

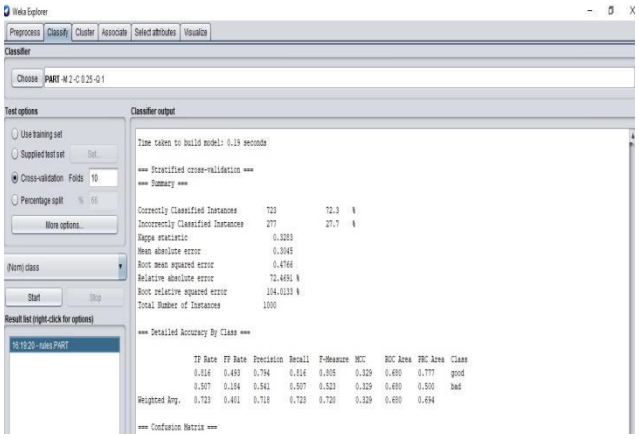

Fig 5 results of PART classifier with respect to the original german_credit dataset

Table 3: Results of PART classifier after applying the four attribute evaluators

Evaluator	Avg.Precision	Avg. Recall	Avg.F-Measure	Avg.Mean-absolute error
CfsSubsetEval (CSE)	0.725	0.736	0.729	0.314
Correlation Attribute Eval (CA)	0.706	0.716	0.714	0.341
GainRatio Attribute evaluation (GA)	0.699	0.717	0.703	0.343
InformationGain Attribute evaluation (IG)	0.722	0.733	0.726	0.324

The results of the PART classifier with respect to the four attribute evaluators are also presented as graph in Fig 6. This graph shows the comparison of four evaluators adapted to the PART classifier with respect average precsion, recall and f-measure.

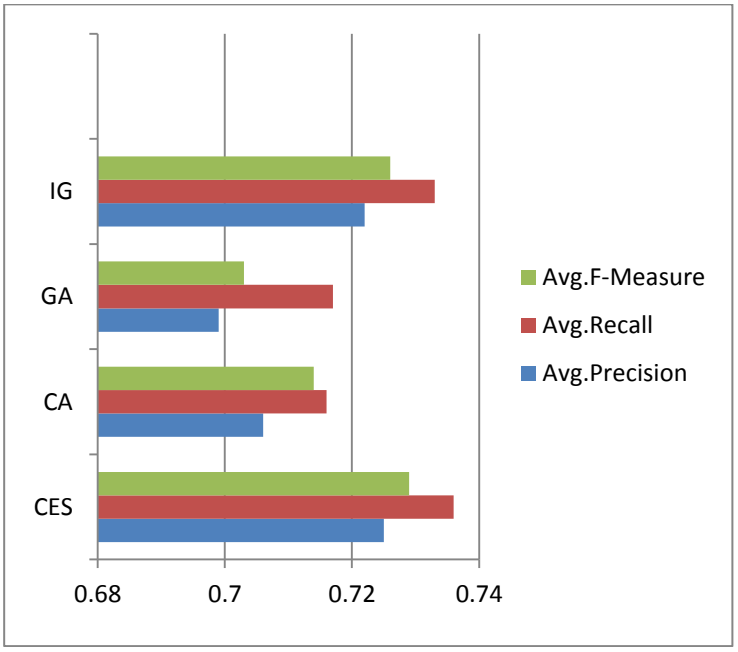

Fig 6: Metrics of PART classifier with respect to four evaluators.

Volume: 09 Issue: 06 Pages: 3640-3644 (2018) ISSN: 0975-0290

V. CONCLUSION

This study compares the four basic attribute selection algorithms with the utilized PART classifier in terms of precision , recall , F-measure and mean absolute error. The study concludes that the performance metrics of the classifier increases with respect to the attribute evaluator. The PART classifier returns good results when it evaluates the attributes by the CfsSubsetEval method for the german_credit dataset.

REFERENCES

[1] Meenatchi V.T, Gnanambal S, et.al, Comparative Study and Analysis of Classification Algorithms through Machine Learning, *International Journal of Computer Engineering and Applications*, 9(1),247-252,2018.

[2] Hany M. Harb1, Malaka A. Moustafa, Selecting optimal subset of features for student performance model, *IJCSI* , 9(5), 2012, 1694-0814

[3] Hwang, Young-Sup,Wrapper-based Feature Selection Using Support Vector Machine, Department of Computer Science and Engineering, Sun Moon University, Asan, Sunmoonro, Korea, *Life Science Journal*,11 (7), 221-70,2014.

[4] Wang Liping, Feature Selection Algorithm Based On Conditional Dynamic Mutual Information, *International Journal Of Smart Sensing and Intelligent Systems*,8(1), 2015.

[5] Qinbao Song, Jingjie Ni and Guangtao Wang, A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data, *IEEE Transactions on Knowledge and Data Engineering*, 25(1), 2013.

[6] Z.Zhao, H.Liu, On Similarity Preserving Feature Selection, *IEEE Transactions on Knowledge and Data Engineering*, 25(3), 2013.

[7] Sunita Beniwal and Jitender Arora, Classification and Feature Selection Techniques in Data Mining, *International Journal of Engineering Research & Technology (IJERT),* 1(6), 2012.

[8] Mital Doshi and Setu K Chaturvedi, Correlation Based Feature Selection (Cfs) Technique To Predict Student Perfromance, *International Journal of Computer Networks & Communications (IJCNC*),6(3),2014.

[9] M. Ramaswami and R. Bhaskaran,A Study on Feature Selection Techniques in Educational Data Mining, *Journal Of Computing*,1(1),December 2009.

[10] K.Sutha and J. Jebamalar Tamilselv, A Review of Feature Selection Algorithms for Data Mining Techniques, *International Journal on Computer Science and Engineering (IJCSE),* 7(6), ,2015.

[11] Gnanambal S and Thangaraj M, A new architectural framework for Rule Based Healthcare System using Semantic Web Technologies, *International Journal of Computers in Healthcare*, Inderscience, 2(1), 2014, 1-14. ISSN: 1755-3202.