Application of Feature Subset Selection methods on Classifiers Comprehensibility for Bio-Medical datasets

Syed Imran Ali[1], Byeong Ho Kang[2] and Sungyoung Lee[1]

[1] Department of Computer Engineering Kyung Hee University Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do, Republic of Korea
{imran.ali,sylee}@oslab.khu.ac.kr

[2] Department of Engineering and Technology, Information and Communication Technology, University of Tasmania, Australia
byeong.Kang@utas.edu.au

Abstract. Feature subset selection is an important data reduction technique. Effects of feature selection on classifier's accuracy are extensively studied yet comprehensibility of the resultant model is given less attention. We show that a weak feature selection method may significantly increase the complexity of a classification model. We also proposed an extendable feature selection methodology based on our preliminary results. Insights from the study can be used for developing clinical decision support systems.

Keywords: Feature Subset Selection, Model Comprehensibility, Data Classification, Data Mining, Clinical Decision Support System

1 INTRODUCTION

Data classification is one of the important tasks in data mining for knowledge acquisition. The main purpose of a classification algorithm is to model relationship between independent features and a response variable. An inferred model constructed by a classification algorithm may produce either a comprehensible model or an incomprehensible model. This paper deals with comprehensible models. Decision tree and rule-based decision list are the two most common used comprehensible models [1]. Aforementioned models are depicted in Fig. 1 and Fig. 2, respectively. Along with predictive accuracy, comprehensibility of a model is also an important characteristic for a classification algorithm in certain domains [1].

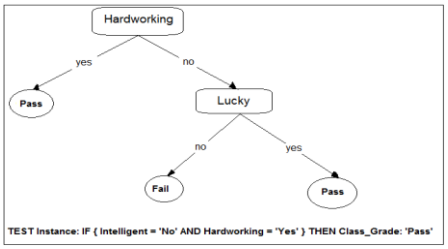

Fig. 1. Decision Tree model

IF <Hardworking=Yes> THEN <Class: Pass>
IF <Hardworking=No> AND <Lucky=No> THEN <Class: Fail>
IF <Hardworking=No> AND <Lucy=Yes> THEN <Class: Pass>
IF <Intelligent=Yes> AND <Lucky=Yes> THEN <Class: Pass>
IF <Intelligent=No> AND <Lucy=No> THEN <Class: Fail>
IF <Lucky=Yes> THEN <Class: Pass>
Default: <Class: Fail>

Test Instance: IF <Intelligent=No> AND <Hardworking=Yes>
THEN <Class: Pass>

Fig. 2. Decision List model

2 RELATED WORK

This section deals with some of the important related studies. Fast Correlation Based Feature Selection (FCBFS) [2] is one of the highly effective filter methods. It accounts for both feature relevancy and redundancy. H. Liu et al. [3] proposed a consistency based feature selection mechanism. We have used genetic algorithm in this method (referred hereafter GA-Consist). Correlation-based Feature subset Selection (CFS) evaluates usefulness of a subset of attributes based inter and intra feature correlation [4]. This study is based on four commonly used comprehensible models namely C4.5, CART, RIPPER and Ant-Miner. Table 1 summarized some of the important studies.

Table 1. Summarized Related Work

Ref	Main Contribution	Limitation(s)
[5]	Surveys FSS methods, proposed categorizing framework and an integrated platform for automatic selection of FSS based on dataset characteristics	No empirical experimentation, classifier comprehensibility not discussed
[6]	Survey FSS methods for intrusion detection systems, empirical experimentation mainly based on predictive accuracy only	Classifier comprehensibility not discussed, no integrated framework proposed
[7]	Proposed a scoring measure to compare results of different FSS methods, Empirical experimentation is performed to contrast the ability of the different FSAs to hit a solution with respect to relevance, irrelevance, redundancy and sample size	No comprehensible framework on the basis of FSS and classifier complexity proposed
[8]	Survey state of the art FSS methods for micro-array datasets, dataset shift and imbalanced ration also discussed, empirical experiment mainly based on predictive accuracy only	Effects of FSS on model size are not studied, no integrating framework proposed
[9]	Proposed a framework for cost based feature , detailed empirical experimentation based on classifier error and total cost of the selected features	Model's complexity is not discussed, no categorizing framework explored

3 Proposed Methodology and Experimentation

Based on the empirical study, we proposed a methodology for selecting a feature selection method based on a number of considerations. The key consider-

ation for this study is the nature of the classification model intended to be produced i.e. true base model or rule based model. Accuracy and complexity of the model are also addressed. For example if a rule based model is intended and complexity of the model is of high consideration then FCBF is recommended otherwise CFS can yield relatively accurate results with comparatively a more complex model. Fig. 3 depicts feature selection methodology in a graphical manner. It is important to note that this is a preliminary study which lays foundation for further studies on the intersection of effective data reduction methods and classifier comprehensibility.

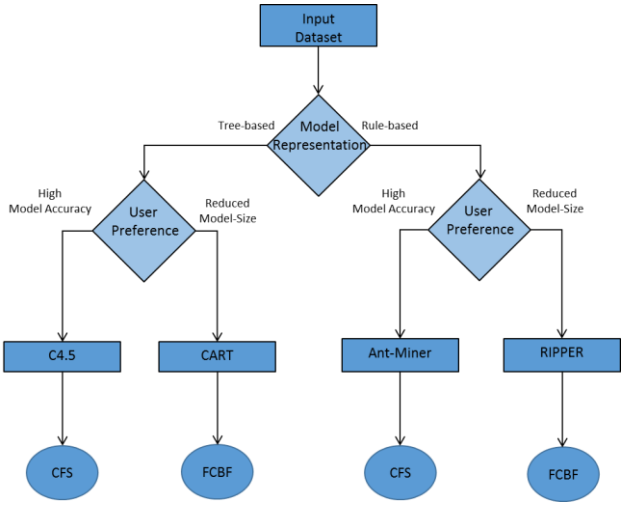

Fig. 3. Feature selection Methodology

The purpose of the proposed methodology is to assist in selecting a feature selection method based on user's requirements. Moreover, only those feature selection methods are retained which enhances some aspect of the classifier e.g. compact model size, predictive accuracy, reduced training/testing time, etc. Since, GA-Consist couldn't provide any extra advantage over either CFS or FCBF therefore it is not depicted in Fig. 3. Datasets used in this study are easily accessible from University of California, Irvine (UCI) machine learning repository [10] and related openly accessible dataset repositories. Table 2 enlists datasets along with basic statistics i.e. number of features, instances and classes in each dataset. It is important to note that all the datasets employed in this study are from the bio-medical domain. All the experiments from Table 3 to Table 6 are performed using 10-fold cross validation. These datasets are of varied complexity i.e. from medium to large dimensionality, high number of instances,

imbalanced feature-instance ratio, etc. Hence, insights gathered from the study can be extended to data driven decision support systems.

Table 2. Datasets characterisitcs

No.	Dataset	Features	Instances	Classes
1	Breast-Cancer	9	286	2
2	Diabetes	8	768	2
3	Heart-c	13	303	2
4	Heart-h	13	294	2
5	Heart-Statlog	13	270	2
6	Hepatitis	19	155	2
7	Hypothyroid	29	3772	2
8	Lymph	18	148	4
9	Primary Tumor	17	339	21
10	Splice	60	3175	3
11	Lung Cancer	56	32	3

In order to construct comprehensible classifiers all the datasets with numeric features are discretized. In this study effect of data discretization on model construction is not studied. Four state of the art classifiers are used i.e. C4.5, CART, RIPPER and Ant-Miner. Three filter methods, FCBF, CFS and GA-Consistency are used for feature subset selection. Detailed experimentation is performed. Due to the page limitation only the summarized results are discussed. As mentioned in Table 3 with no feature selection C4.5 achieved an average accuracy of 76.47%. Average tree size and number of leaves are 42.63 and 31.81, respectively. In case of feature selection it can be observed that all the feature selection methods achieved a reduced dataset. FCBF selected the smallest number of features i.e. on average 7.63 features were retained, while the tree size was also reduced significantly. Average accuracy of C4.5 is slightly reduced with FCBF as compared to no feature selection. CFS also achieved smaller tree size as compared to without feature selection. Moreover, its accuracy is slightly better than the feature selection methods considered in the study. Case of GA-Consist is interesting. It could not perform well for classification accuracy although it retained only half of the feature set on average which is a lot of feature reduction. Tree size in case of GA-Consist is far more complex than with any of the other feature selection algorithms or with no feature selection. It can be easily observed that due to the failure in selecting useful features it not only affects the accuracy of the classifier but the model size also complicates more so than using all the features. It is an important observation

which allows for more study on the effects of feature selection methods on the resultant model's complexity.

Table 3. Experimental results for C4.5

	Average Features	Average Accuracy	Average Tree Size	No. of Leaves
Total	23.18	76.47	42.63	31.81
FCBF	7.63	74.5	33.9	26.36
CFS	9.45	75.97	38	27.81
GA-Consist	12	71.65	66.76	49.07

Table 4. Experimental results for CART

	Average Features	Average Accuracy	Average Tree Size	No. of Leaves
Total	23.18	72.89	16.09	8.54
FCBF	7.63	74.86	16.63	8.81
CFS	9.45	73.40455	16.63	8.81
GA-Consist	12	70.9	19	10

CART classifier achieved an average accuracy of 72.89% with no feature selection. Resultant tree size and number of leaves are 16.09 and 8.54, respectively. CART could achieve lower accuracy as compared to C4.5 with a much reduced tree size on average. FCBF achieved highest accuracy on CART yet couldn't much improve on the model size. GA-Consist resulted in lower accuracy on average and a comparatively large model size.

For rule-based classifiers we have opted for a popular model complexity metric i.e. number of rules and conditions per rule [1]. RIPPER achieves an average accuracy of 72.41% with no feature selection. On average 4.81 rules were created with a 1.96 conditions-to-rule ratio. Moreover, FCBF achieved slightly higher accuracy than CFS. Model size of FCBF was also slightly larger than with no feature selection, as is the case with other two feature selection methods. Hence, it can be observed that in case of RIPPER classifier feature selection methods couldn't improve on the model size of the algorithm. So here we can observe the trade-off between classifier's accuracy and its model size. If a user has a preference for the former case then FCBF would be preferred while no feature selection method may be preferred for the latter case.

Table 5. Experimental results for RIPPER

	Features	Accuracy	Rules	Conditions	Conditions/ Rules
Total	23.18	72.41	4.81	9.45	1.96
FCBF	7.63	75.1	4.9	10	2.03
CFS	9.45	75	6	13.75	2.29
GA-Consist	12	71.52	8.16	20	2.44

Table 6. Experimental results for Ant-Miner

	Features	Accuracy	Rules	Conditions	Conditions/ Rules
Total	23.18	74.64	8.04	19.21	2.38
FCBF	7.63	79.13	7.01	9.87	1.4
CFS	9.45	80.14	7.62	12.93	1.69
GA-Consist	12	72.35	8.25	16.66	2.01

Table 6 mentions results for Ant-miner classifier. On average Ant-miner algorithm achieves higher accuracy than RIPPER. Moreover, average model size of Ant-miner is comparatively larger than that of RIPPER. CFS achieved the highest accuracy, with a slightly larger model size then FCBF. Although GA-Consist couldn't achieve higher accuracy but it did achieve a lower model size. So we can infer that the effects of a feature selection method on the model size vary from one classifier to another. Hence, effects of feature selection on comprehensibility of the classifier are more subtle. Since Ant-miner is a population-

based stochastic algorithm [11] it has incurred the highest training time of all the classifiers considered in this study.

4 CONCLUSION AND FUTURE WORK

In this study we evaluated effects of feature selection methods on comprehensibility of the classifiers. Classifier comprehensibility has received relatively less attention while selecting for an appropriate feature selection method. We have shown that different feature selection methods have a varied effect on the comprehensibility of classifiers.

Acknowledgments. This work was supported by the Industrial Core Technology Development Program (10049079 , Develop of mining core technology exploiting personal big data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea)" and This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) NRF-2014R1A2A2A01003914.

REFERENCES

[1] Freitas, A.A., 2014. Comprehensible classification models: a position paper. ACM SIGKDD explorations newsletter, 15(1), pp.1-10.

[2] Yu, Lei, and Huan Liu. "Feature selection for high-dimensional data: A fast correlation-based filter solution." In ICML, vol. 3, pp. 856-863. 2003.

[3] Liu, Huan, and Rudy Setiono. "A probabilistic approach to feature selection-a filter solution." In ICML, vol. 96, pp. 319-327. 1996.

[4] Hall, Mark A. Correlation-based feature selection for machine learning. Diss. The University of Waikato, 1999.

[5] Liu, H. and Yu, L., 2005. Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on knowledge and data engineering, 17(4), pp.491-502.

[6] Chen, You, et al. "Survey and taxonomy of feature selection algorithms in intrusion detection system." Information security and cryptology. Springer Berlin Heidelberg, 2006.

[7] Belanche, L.A. and González, F.F., 2011. Review and evaluation of feature selection algorithms in synthetic problems. arXiv preprint arXiv:1101.2320.

[8] Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J.M. and Herrera, F., 2014. A review of microarray datasets and applied feature selection methods. Information Sciences, 282, pp.111-135.

[9] Bolón-Canedo, V., Porto-Díaz, I., Sánchez-Maroño, N. and Alonso-Betanzos, A., 2014. A framework for cost-based feature selection. Pattern Recognition, 47(7), pp.2481-2489.

[10] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[11] Parpinelli, R.S., Lopes, H.S. and Freitas, A.A., 2002. Data mining with an ant colony optimization algorithm. IEEE transactions on evolutionary computation, 6(4), pp.321-332.