An Evaluation of Feature Selection Methods on Multi-Class Imbalance and High Dimensionality Shape-Based Leaf Image Features

Mohd Shamrie Sainin[1], Rayner Alfred[2], Faudziah Ahmad[1], Mohamed A.M Lammasha[1]
[1]Computational Intelligence Research Group, School of Computing, College of Arts and Sciences, Universiti Utara Malaysia.
[2]Faculty of Computing and Informatics, Universiti Malaysia Sabah.
shamrie@uum.edu.my

***Abstract*—Multi-class imbalance shape-based leaf image features requires feature subset that appropriately represent the leaf shape. Multi-class imbalance data is a type of data classification problem in which some data classes is highly underrepresented compared to others. This occurs when at least one data class is represented by just a few numbers of training samples known as the minority class compared to other classes that make up the majority class. To address this issue in shape-based leaf image feature extraction, this paper discusses the evaluation of several methods available in Weka and a wrapper-based genetic algorithm feature selection.**

***Index Terms*—Feature Selection; Multiclass Imbalance; High Dimensionality; Leaf.**

I. Introduction

Class imbalance data is a type of data classification problem in which some data classes are highly underrepresented. This occurs when at least one of the data class is represented by just few numbers of training samples known as the minority class [1]. Hence, the problem to learn in such conditions constitutes most biases particular to several learning algorithms which are the most significant in some real applications such as biological data analysis, image classification, text classification, and web page classification.

In the earlier time, several approaches have been explored to provide solution to this pervasive problem. The most significant point is that these methods are just based on two-class imbalanced problems. Meanwhile, in practice, many problem domains possess more than two classes with uneven distributions, which a typical example such as in protein fold classification and the weld flaw classification [2].

Multiclass imbalance problems present new challenges that are not associated with two-class problems. Although that two-class problem can be extended to multi-class problem, however, it was presented that the multi-class imbalance problems pose the challenges that are not covered by the two-class problems [3]. Thus, further investigation is necessary to examine multiclass imbalance for specific problems.

Methods for imbalance problem can be categorized in two groups based on their approaches, namely data-level (also known as sampling) and algorithm-level [4]. Data-level methods are concerned about how the data are presented to the classifier to address the imbalance problem. There are two methods that are associated with data-level methods which are column-based (feature selection) and row-based (e.g. sampling).

Feature selection (also known as attribute subset selection or attribute reduction) is an important research issue in data mining and machine learning, and can be viewed as part of data pre-processing techniques (selecting the subsets of the available features [5]).

The study reported in this paper considers feature selection to address the imbalanced multiclass problems, mainly to find the subset of relevant features and to improve the prediction accuracy.

There are many feature selection methods that specifically implemented to reduce the dimensionality of features in the data. Many researchers have recently put more focus on feature selection. With the quick progress of computers and database innovation advancements, datasets with a large number of variables or components are currently omnipresent in recognition of pattern, data mining, and machine learning [6]. Feature selection can be addressed in three general schemes; filter methods, wrapper methods and embedded methods [5]. The schemes for feature selection differ significantly on how the search operates on a given feature. Filter methods look at the problem of feature selection as an independent process from the model selection (i.e. inductive generalization is not involved in feature selection process). In contrast, the wrapper methods associate the hypotheses search with the inductive classifiers to get the feedback whether the model selection is good. In this method, various combinations of feature subsets are generated and evaluated in order to improve the classification performance. On the other hand, embedded methods search for an optimal subset and are designed internally to the classifier construction.

Feature selection approaches have been applied in various studies such as dendrite cell algorithm, dynamic software quality attributes selection, UCI benchmark dataset, intrusion detection, association rule, and imbalance learning. Among the method presented in these studies are comparisons of feature selection methods (e.g. Information Gain, Gain Ratio, etc.) [7], wrapper-based genetic algorithm [8, 9], filter-wrapper with feature ranking [10], mixed ant colony optimization [11], and filter-based comparison of symmetrical tau and mutual information [12].

In imbalance learning, specific feature selection study has been presented using ensemble-based wrapper which utilizing hybrid sampling and trainable base classifiers that consists of Random Forest, 3-Nearest Neighbors, 7-Nearest Neighbors, LogReg, Multi-Layer Perceptron and ADTree [13]. The finding from the study reveals that by incorporating the ensemble with feature selection and multiple sampling shows that, features selection approaches using ensemble-based wrapper are significantly better than using single classifier.

The domain of Machine learning and Data mining researches are greatly faced with the problem of multiclass imbalance. This is most significant in some of the real world applications. Although class imbalanced problem has been extensively investigated. However, the issue of high dimensionality in data still remain unsolved, whereas, high dimensionality is a common feature of class imbalance problem. In a study centered on Malaysian medicinal leaf identification [14], leaf shape features generate enormous possibilities (high dimensional data) for leaf species. In this kind of identification, though, high accuracy may be recorded by a classifier in identifying the dominant leaf features (majority class) but there is greater tendency for the same classifier to record low or very low performance in identifying the non-dominant features.

Herdiyeni and Santoni [15] show an identification performance of 72.16% for herb leaf features. In the same vein, Prasvita and Herdiyeni [16] presents 56.33%, which is a result obtained for 30 species of Indonesian medicinal plants having 48 different image features. Similarly, leaf identification by Sainin, Ahmad and Alfred [14] has achieved 70% accuracy on high dimensional shape-based features using ensemble classifier. It is obvious in these studies that despite the efforts made in extracting the necessary features for identification of medicinal leaf, optimum classification feature is still unable to be achieved by most of the classifiers.

Therefore, in this study, we aim to investigate the performance of different feature selection algorithms for shape-based leaf features and evaluate their effectiveness in representing the relevant shape features. We use feature selection approaches available in Weka [17] according to evaluator and search methods. Evaluator that directly reduce the features are CfsSubsetEval, ConsistencySubsetEval, FilteredSubsetEval and WrapperSubsetEval. Search methods are including BestFirst, GeneticSearch, and GreedyStepwise and LinearForwardSelection. In addition, wrapper-based feature selection based on genetic algorithm[8] is also investigated. The assesment of these methods is according to two measurement metrics: accuracy and F-measure. This paper is organized as follows. Section 2 introduces the overview of the shape-based leaf features. Then, the related works in feature selection and methods used in this study is discussed in Section III. Experiment setup in Section IV and Section V discusses the finding of the study. The conculsion of this work is presented in the final section.

II. SHAPE-BASED LEAF FEATURES

Shape-based is the one of the popular approaches for feature extraction as it provides rich information for classification. The earliest work in leaf shape-based automated identification on specific leaf is started by Heymans and Kuti [18] which involves extracting the shape of the leaf (represented as grid) and using neural network for identification purposes. Since then, other techniques based on shape features were presented such as centroid contour distance [19, 20], inner-distance shape context approach [21], and Moving Median Center Hypersphere [22]. Several shape-based features were compared for their recognition, namely geometric features, moment invariants, Zernike moments and Polar Fourier Transform (PFT), where PFT gives the best result. In recent works, Patil and Manza [23] presented various shape-based feature extraction methods based on geometric and morphological features.

In other method, shape features such as region, aspect ratio, area and perimeter were used to represent the leaf shape in Singh and Bhamrah [24]. Their result based on Neural Network classification is 98%. Using the same classifier method, shape method with Moments-Invariant (M-I) model and Centroid-Radii (C-R) model that was applied to 180 images with three classes provides 88.9% and 100% respectively.

Most of the studies in shape based leaf features consider balanced multiclass data. However, in real world application, leaf collection or sampling may not create balanced data among leaf species. This is due to available sampling where some plants may have few leaves that can be used as sample. This condition poses a new problem in the identification called multiclass imbalance data. In relation to this problem, leaf shape feature extraction addresses the imbalance problem and high dimensionality [25] shows low performance using five species and 65 leaves. The best accuracy is 50% using SMO in Weka and 65% using ensemble classifier. Therefore, this paper is investigating the effect of feature selection methods on shape-based leaf feature.

III. FEATURE SELECTION METHODS

As mentioned in previous section of this paper, Weka's feature selection evaluators and search methods are investigated for their effect on the multiclass imbalance classification performance. Another wrapper-based feature selection using genetic algorithm is also examined.

A. CfsSubsetEval, ConsistencySubsetEval and FilteredSubsetEval

These evaluators are in the type of filter-based feature selection. CfsSubsetEval or Correlation-based feature selection method (CFS) is concern with the hypothesis which contain features that are highly correlated with the class, but has no correlation with each other [26]. In that sense, each feature is the test that measure traits related to the class using merit evaluation [27]. It then will compute the correlation between attributes by first, applying the discretization and followed by the symmetrical uncertainty measure. In the study, CFS is proven to be comparable to wrapper feature selection method, but better on small datasets and overall running time [27, 28].

ConsistencySubsetEval (CSE) is based on probabilistic approach to feature selection that is claimed to be simple and fast feature selection algorithm, thus guaranteed to find the optimal given the suitable resources [29]. The probabilistic approach called Las Vegas Algorithm (LVF) makes probabilistic choices as guide for the search of feature subset. In the experiment of this filter-based feature selection, it produces minimum features for the tested datasets with promising error rates.

An analytical comparison on filter based feature selection has been conducted on CFS and CSE using decision tree classifier for accuracy measurement [30], where CFS provides less feature subset most of the time but CSE with BestFirstSearch strategy has higher perforamance.

FilteredSubsetEval (FSE) is simply a filter-based feature selection which available in Weka that running an arbitrary subset evaluator on the training data and produce the best feature subset [31].

B. *WrapperSubsetEval and WrapperGA*

WrapperSubsetEval [32] in Weka is a feature selection method that using an induction algorithm as a blackbox (evaluator) for feature subset, where accuracy estimation technique is applied to measure how good is the features. In the study, the method is shown to improve significantly for some datasets with two induction algorithm namely decision tree and Naïve Bayes.

Wrapper-based selection method using genetic algorithm (GA) used in this study is implemented according to the genetic based wrapper feature selection (WrapperGA) approach using nearest neighbor distance matrix [8]. In this method, a supervised nearest neighbor distance matrix (NNDM) which produces the nearest neighbor matrix during training. The NNDM is also applying the loss function in order to group similar class label to each instance. In order to select the feature subset, The GA based wrapper feature selection is adopted to optimize the possible combination of certain number of attributes that best describe the dataset, where fitness function is based on the best information gain score and using Naïve Bayes and kNN as the ensemble classifier (called DECIML) performance evaluator. The method is shown to produce higher classification but with higher number of feature subset. Thus, the method is investigated in multiclass imbalance and high dimensionality leaf shape data.

IV. Experimental Setup

The dataset of Malaysian medicinal leaf images [14] is acquired to follow closely the original dataset in order to compare the classification performance of the best feature selection outcome. Species of the leaves are presented in Table 1. The dataset is available upon request to the corresponding author.

The shape-based features from the leaf images are acquired by using shapes represented as angles of each point specified in the leaf. A full-leaf shape produces about 624 angles (contour points along the leaf boundary, represented as sinus and cosines) which then become the attributes. Table 2 shows the description of the experimental data. Although the dataset looks small, but it dubs a high dimensionality, where the obtained feature space dimension is 45*624 for training set and 20*624 for test set. This leads to face the high dimensionality problem with multiclass imbalance, whereby, this paper suggests using the feature selection methods and comparing their performance.

The experiment uses three filter-based feature selection methods (CFS, CSE, and FSE), each using search methods (BestFirst (BF), GeneticSearch(GS), GreedyStepwise(GSW) and LinearForwardSelection(LF)) and full training set approach. In FSE, the filter used is RESAMPLE with biasToUniformClass=1.0 and sampleSizePercent=150%, and also SMOTE. The setting is to make sure that the instances in each class is balanced. Wrapper-based feature selection (WrapperSubsetEval) in Weka is implemeted using Naïve Bayes as the induction algorithm. Another experiment is the WrapperGA8 with parameters (crossover rate=0.7, mutation rate=0.01 and maximum generation=100). The classifiers that were used to evaluate the performance of the feature selection methods in Weka are Naïve Bayes (NB), J48, and Random Forest (RF).

Table 1
Experiment Dataset

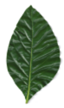
Class	Example	Name	Train	Test
1		Cemumar	11	4
2		Kapal Terbang	12	4
3		Kemumur Itik	11	4
4		Lakom	5	4
5		Mengkudu	6	4
	Total		45	20

Table 2
Dataset description

Description	Value #
#Examples	65
#Attributes	624
#Training	45
#Testing	20
#Majority	12
#Minority	5

Evaluation metric that was observed in each experiment is the F-measure, which is normally used to measure the true positive rate as well as the accuracy of positive prediction (in multiclass imbalance problem). The highest value in F-measure indicates that the performance of the feature selection methods given by the classifier is the best result.

V. Findings

This section discusses the evaluation of the Weka's feature selection performance and the wrapper method over the data. First, the classification performance (accuracy) of the classifiers when using all features is listed in Table 3, where RF provides higher classification accuracy (70%). Then, the number of feature selected by the evaluator and search methods is presented in Table 4.

Table 3
Accuracy using all features

	NB	J48	RF
Accuracy (%)	50	60	70

Table 4
Number of feature selected using the evaluator and search method.

	BF	GS	GSW	LF
CFS	9	229	9	6
CSE	8	266	6	5
FSE+Resample	20	176	20	8
FSE+SMOTE	14	261	14	3
Wrapper+NB	8	291	3	11
Wrapper GA	-	384	-	-

Based on Table 4, CFS and CSE produce less than 10 attributes (out of 624) when using BF, GSW and LF search methods. It can be seen that LF search method gives minimum number of features consistently in the experiments, where FSE+SMOTE output the minimum features (3 features). Further tests were carried out to compare the performance of selected features using three classifiers as shown in Table 5. Performance using all features shows that it is better than most of feature selection methods. According to the results, filter-based feature selection methods almost perform similar given by the three classifiers. It can be seen that FSE+Resample performs better in every tested classifier, where GS is the best search method when J48 and Random Forest is used as the classifiers. However, BF performs averagely better in all feature selection methods when Naïve Bayes is used as classifier. Interestingly, despite that FSE+Resample perform better in average, only NB evaluates the feature selection methods with 70% accuracy using BF or GSW as search technique.

In relation to small feature subset selection, LF is surprisingly comes in second place for the average performance. In fact, although not the most minimum number of features, LF combined with Wrapper+NB provides the highest classification accuracy of 75%. This shows that the wrapper-based feature selection method has successfully selected the best feature subset (11 features) and evaluated by Random Forest (using default settings).

Table 5
Performance of the feature selection methods using NB, J48 and Random Forest

Methods	NB/Search BF	GS	GSW	LF	Avg.
CFS	50	40	50	55	48.75
CSE	45	65	45	20	43.75
FSE+Resample	**70**	40	**70**	55	**58.75**
FSE+SMOTE	50	50	50	40	47.5
Wrapper+NB	45	50	35	45	32.5
Average	**52**	49	50	34	
Methods	**J48/Search BF**	**GS**	**GSW**	**LF**	**Avg.**
CFS	35	50	35	45	41.25
CSE	25	55	20	40	35
FSE+Resample	50	60	50	50	**52.5**
FSE+SMOTE	35	60	35	50	45
Wrapper+NB	60	50	30	45	46.25
Average	41	**55**	34	46	
Methods	**Random Forest/Search BF**	**GS**	**GSW**	**LF**	**Avg.**
CFS	65	60	65	40	57.5
CSE	50	65	35	45	48.75
FSE+Resample	50	65	50	70	**58.75**
FSE+SMOTE	50	50	50	60	52.5
Wrapper+NB	55	70	30	**75**	57.5
Average	54	**62**	46	58	

Comparing the classification performance between all features and selected features (which is very small), the feature selection effect is notable where it can represent the dataset significantly using small number of features. The performance is similar or even better as shown by the FSE+Resample and Wrapper +NB. Thus, it is proven that feature selection can improve the classification and in the also reduce the running time than using all features. Taking the methods with higher classification rates from Table 5, the detailed performance (F-measure) on the class labels for each method is shown in Table 6. The F-measure values indicate that although the accuracies of some methods are similar, however the effects of feature selection to the imbalance data are varies. It is proven that when the dataset has imbalance problem, high accuracy is actually a poor choice for model evaluation as it just rely on majority class. The 1st row in the table illustrates this problem, where the majority class gets high F-measure while the minority class (Lakom) gives low F-measure value.

Two of the methods show that the F-measure values are almost balance. In this case, the FSE+Resample with BF and GSW have better performance distribution except for class Kapal Terbang, where this class is supposed to be the majority class. FSE+Resample with LF search method higher F-measure on minority class but in turn, gets lower value in majority class, thus, the weighted average accuracy based on F-measure is actually lower (0.69) than the percentage accuracy (70%). Unfortunately, wrapper-based feature selection namely Wrapper+NB with GS has the similar F-measure values when all features are used. An interesting observation on Wrapper+NB with LF that this method provides the best accuracy, however the F-measure values are not seen promising compared to FSE+Resample with BF or GSW, where the majority and minority class gives low F-measure values. Higher average F-measure for Wrapper+NB is achieved due to increased performance in class 'Kemumur Itik'.

Another wrapper-based feature selection using GA and ensemble classifier is investigated for its performance. The fact that GA is an optimization function, it is expected that feature subset combination is optimized so that the classification accuracy will be high. Although that this method gives the highest number of feature selection (384), however, the classification accuracy is 80% with weighted average F-measure is 0.85.

Table 6
Performance (F-measure) of the best feature selection methods on each class (*best value)

	Search	Classifier	Cemumar	Kapal Terbang	Kemumur Itik	Lakom	Mengkudu	Average
All Features	-	RF	0.33	0.80	**0.73***	0.40	1.00	0.65
FSE+Resample	BF	NB	**0.89***	0.44	0.50	0.67	1.00	0.70
FSE+Resample	GSW	NB	**0.89***	0.44	0.50	0.67	1.00	0.70
FSE+Resample	LF	RF	0.67	0.33	0.55	**0.89***	1.00	0.69
Wrapper+NB	GS	RF	0.33	**0.89***	0.67	0.40	1.00	0.66
Wrapper+NB	**LF**	**RF**	**0.89***	0.40	**0.73***	0.57	1.00	**0.72**

VI. CONCLUSIONS

The effective technique for feature selection on multiclass imbalance shape-based data is investigated in this study which consists of five feature selection. From the experiments, there are two important points in order to maintain or increase the classification accuracy while balanced performance among the classes is achieved. First, filter-based method FSE+Resample with BF or GSW search technique is comparable to wrapper-based method (specifically the Wrapper+NB) in terms of performance but faster running time. Second, if number of features is concern, filter-based method is preferable due to small number of features is guaranteed to be produced. However, if the optimized features are required, then wrapper-based GA feature selection can be used but with the cost of running time.

ACKNOWLEDGEMENT

This work was supported in part by FRGS under Grant No. 13142 (2014).

REFERENCES

[1] H. Pant, R. Srivastava. A Survey On Feature Selection Methods for Imbalanced Datasets, International Journal of Computer Engineering & Applications, IX(II) (2015) 197–204.

[2] [2] S. Wang, X. Yao. Multiclass Imbalance Problems: Analysis and Potential Solutions, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2(4)(2012) 1119–1130.

[3] [3] B. Lerteerawong, M. Athimethphat. An Empirical Study of Multiclass Classification with Class Imbalance Problems, International Conference on Business and Information, (2011).

[4] [4] C. Lemnaru, R. Potolea. Imbalanced Classification Problems: Systematic Study, Issues and Best Practices, Lecture Notes Bussiness Intelligence, Springer-Verlag. 102 (2011) 35-50.

[5] [5] L. Ladha, T. Deepa. Feature selection methods and algorithms, International Journal Computational Science Engineering, 3 (2011) 1787–1797.

[6] [6] M. Bekkar, D. T. A. Alitouche. Imbalanced Data Learning Approaches Review, International Journal of Data Mining & Knowledge Management Process, 3(4)(2013) 15–33.

[7] [7] M. F. M. Mohsin, A. R. Hamdan, A. A. Bakar. An Evaluation of Feature Selection Technique for Dendrite Cell Algorithm, in Proceeding of International Conference on IT Convergence and Security, (2014) 1–5.

[8] [8] M. S. Sainin, R. Alfred. A genetic based wrapper feature selection approach using Nearest Neighbour Distance Matrix, in 3rd Conference on Data Mining and Optimization, (2011) 237–242.

[9] [9] M. Barati, A. Abdullah, R. Mahmod, N. Mustapha, N. I. Udzir. Feature Selection for IDS in Encrypted traffic using Genetic Algorithm, Proceedings of the 4th International Conference on Computing and Informatics, (2013) 279–285.

[10] [10] S. S. Kamaruddin, J. Yahaya, A. Deraman, R. Ahmad. Filter - Wrapper based Feature Ranking Technique for Dynamic Software Quality Attributes, Knowledge Management International Conference (KMICe), (2012) 593–597.

[11] [11] H. B. Alwan, K. R. Ku-Mahamud. Mixed variable ant colony optimization technique for feature subset selection and model selection, Proceedings of the 4th International Conference on Computing and Informatics, (2013) 25–31.

[12] [12] I. N. M. Shaharanee, J. Jamil. Feature selection and rule removal for frequent association rule based classification, Proceedings of the 4th International Conference on Computing and Informatics, (2013) 377–382.

[13] [13] P. Yang, W. Liu, B. B. Zhou, S. Chawla, A. Y. Zomaya. Ensemble-Based Wrapper Methods for Feature Selection and Class Imbalance Learning, Advanced Knowledge Discovery in Data Mining, 7818 (2013) 544–555.

[14] [14] M. S. Sainin, F. Ahmad, and R. Alfred. Improving the identification and classification of Malaysian medicinal leaf images using ensemble method, Proceeding of the First International Conference on ICT for Transformation, (2016) 1–6.

[15] [15] Y. Herdiyeni, M. M. Santoni. Combination of Morphological, Local Binary Pattern Variance and Color Moments Features for Indonesian Medicinal Plants Identification, International Conference on Advanced Computer Science and Informations, (2012) 255–259.

[16] [16] D. S. Prasvita, Y. Herdiyeni. Medleaf Mobile Application for Medicinal Plant Identification Based on Leaf Image, International Journal on Advanced Science, Engineering and Information Technology, 2(2)(2013) 5–8.

[17] [17] I. Witten, E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann. (2000).

[18] [18] B. C. Heymans, J. P. Onema, J. O. Kuti. A neural network for opuntia leaf-form recognition, IEEE International Joint Conference on Neural Networks, 3 (1991) 2116–2121.

[19] [19] F. Hong, C. Zheru, F. Dagan, S. Jiatao. Machine learning techniques for ontology-based leaf classification, 8th Control, Automation, Robotics and Vision Conference, 1 (2004) 681–686.

[20] [20] Z. Wang, Z. Chi, D. Feng. Shape based leaf image retrieval, IEE Proc. Vision, Image Signal Process., 150(1)(2003) 34–43.

[21] [21] M. V Joshi, V. Kumar, and R. C. Agarwal. (2001) “Evaluating boosting algorithms to classify rare cases: comparison and improvements,” First IEEE International Conference on Data Mining, 257–264.

[22] [22] J.-X. Du, X.-F. Wang, G.-J. Zhang. Leaf shape based plant species recognition, Journal Applied Mathematics and Computation, 185(2) (2007) 883–893.

[23] [23] V. R. Patil, R. R. Manza. A Method of Feature Extraction from Leaf Architecture, Int. J. of Advanced Research in Computer Science and Software Engineering, 5(7) (2015) 1025–1029.

[24] [24] S. Singh, M. S. Bhamrah. Leaf Identification Using Feature Extraction and Neural Network, IOSR Journal of Electronics and Communication Engineering, 5(I) (2015) 134–140.

[25] [25] M. S. Sainin, T. K. Ghazali, R. Alfred. Malaysian Medicinal Plant Leaf Shape Identification and Classification, Knowledge Management International Conference and Exhibition, (2014) 578–583.

[26] [26] M. A. Hall. Correlation-based Feature Subset Selection for Machine Learning, Univerisiy of Waikato, (1999).

[27] [27] M. A. Hall. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning, Proceedings of the Seventeenth International Conference on Machine Learning, (2000) 359–366.

[28] [28] M. A. Hall, L. A. Smith. Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper, Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference, (1999).

[29] [29] H. Liu, R. Setiono. A probabilistic approach to feature selection - A filter solution, 13th International Conference on Machine Learning, (1996) 319–327.

[30] [30] A. R. Onik, N. F. Haq, L. Alam. An Analytical Comparison on Filter Feature Extraction Method in Data Mining using J48 Classifier, International Journal of Computing Application, 124(13) (2015) 1–8.

[31] [31] G. Cuaya, A. Muñoz-Meléndez, E. F. Morales, A Minority Class Feature Selection Method, Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 7042(2011) 417–424.

[32] [32] R. Kohavi, G. H. John. Wrappers for feature subset selection, Artificial Intelligence, 97(1-2) (1997) 273–324.