Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

SELECTION OF CLASSIFICATION MODELS FROM REPOSITORY OF MODEL FOR WATER QUALITY DATASET

M. Makhtar[1,*], S. Y. Muhammad[1], A. Rozaimee[1], S. N. W. Shamsuddin[1], N. Zamri[1] and M. K. A. Kamarudin[2]

[1]Faculty of Informatics and Computing, Universiti Sultan ZainalAbidin, Tembila Campus, 22200 Besut, Terengganu, Malaysia

[2]East Coast Environmental Research Institute (ESERI), Universiti Sultan ZainalAbidin, Gong Badak Campus, 21300 Kuala Terengganu, Terengganu, Malaysia

Published online: 10 November 2017

ABSTRACT

This paper proposes a new technique, Model Selection Technique (MST) for selection and ranking of models from the repository of models by combining three performance measures (Acc, TPR and TNR). This technique provides weightage to each performance measure to find the most suitable model from the repository of models. A number of classification models have been generated to classify water quality using the most significant features and classifiers such as J48, JRip and BayesNet. To validate this technique proposed, the water quality dataset of Kinta River was used in this research. The results demonstrate that the Function classifier is the optimal model with the most outstanding accuracy of 97.02%, TPR = 0.96 and TNR = 0.98. In conclusion, MST is able to find the most relevant model from the repository of models by using weights in classifying the water quality dataset.

Keywords: selection of models; water quality; classification model; models repository.

Author Correspondence, e-mail: mokhairi@unisza.edu.my

doi: http://dx.doi.org/10.4314/jfas.v9i6s.56

1. INTRODUCTION

In a data mining classification model, classifiers are measuring both the performance of corrected classified instances and uncorrected classified instances to evaluate the result. This

method has some shortcomings in the way that two or more classifiers may have the same accuracy. On the other hand, it may result in unbalanced datasets. The dataset is unbalanced when the minority class is smaller than the majority class. To fathom these issues, the idea of introducing some other performance measures such as true positive rate (TPR), true negative rate (TNR), false negative rate (FNR) and false positive rate (FPR) in evaluating the result of a classification model is appreciable. On the one hand, it provides a detailed result of the performance of a classifier to avoid the similarity in accuracy of two or more classifiers.

Classifiers are measured on the performance of classified instances to assess the result. This method has a disadvantage in that it may result in poor accuracy. The primary objective of this exploration is to select the suitable classification model for the classification of water quality. The determination of the best classification algorithm for a given dataset is an extremely broad issue, happening every time one needs to choose a classifier to solve a true issue [1]. We may work to distinguish the most exact classifier in some settings, trying to find out the accuracy over yield scores or some combination of yield scores and features considered in the text analysis [2].

Generally, the classifier classifies dataset based on algorithms. This method during selection of a model from the pools of models has some shortcomings, where two or more models may have the same accuracy. To tackle this issue, the two performance measures (TPR and TNR) will be considered in addition to accuracy. This method will provide a detailed result where the model which has the highest accuracy (Acc), true positive rate (TPR), true negative rate (TNR) is the target model. There is a need to come up with a new approach to equate the classifier performance through a variety of measures instead of the normal concentration on the rate (percentage) of precise classifications [3].

There are numerous articles composed by researchers on the selection of a model from a repository of models in machine learning techniques. The majority of the articles have distinctive experimental strategies and a diverse range of study from others.

The main measures utilized for accuracy evaluation, from a particular classification outlook. They studied the case where one wishes to compare diverse classification algorithms and testing them on the agreed data sample, in order to conclude which one will be the ideal (best) on the sampled population [4].

It is a challenging job to discover the optimal classifier by following this process. If we are interested in applying these algorithms to a particular problem, then we need to consider which algorithm is more appropriate for which issue. The suitability test should be possible from guidelines with the assistance of dataset qualities joined with knowledge about how the

diverse algorithms achieve these datasets [3].

The geometric complexity issue in the model selection is investigated. It projected a model selection criterion taken into their account of analysis, and utilized straightforward experiments to verify their technique [5]. Furthermore, through the data geometric method, it obtainedthe natural property of a statistical manifold. It proposed a new model named multistage selection-fusion model (MSF) to improve the simplification of the selection combination of classifiers. The technique combines the selection and fusion of the classifier yields at several layers. The new technique provides simplified performance related to the combination selected and individual best via the single-layer selection model [6].

These applications range from bioinformatics, telecommunications management, speech recognition, text classification to detection of oil spills in satellite images [7].

From the studies, there is a need to select the most relevant models from the repository of model by combining other performance measures because of imbalance dataset scenario and diversity of models generated.

2. METHODOLOGY

The methodology obeys the knowledge discovery in database (KDD) processes. It starts with the data preprocessing which involves the process of discarding or throwing away the unwanted data, converting the non-numeric data to numeric data and filling the gap of the missing data. The feature selection comprises of the feature selection algorithm, which is the determination to distinguish the most significant features or attributes, which are utilized to produce predictive models on a preparation or training datasets.

Selection of data mining task provides a way to choose the data mining task that will be executed in the system such as data mining tasks are classification, regression, etc. Selection of data mining classifier is the step whereby the classifiers would be chosen and it would be used to execute the data mining task. Such classifiers are Tree classifier, Rules classifier etc. Data mining is where the patterns of the executed data mining will be obtained after the executions. Performance measure is the evaluation or assessing the result done by the classifier to provide accuracy.

Basically, we took advantage of seven classifiers with their algorithms to find the optimal model. The classifiers are Rules classifier using JRip algorithm, Tree classifier using J48 graft algorithm, Function classifier using Simple logistic algorithm, Bayes classifier using Bayes Net algorithm, Lazy classifier using LWL algorithm, Meta classifier using Grading algorithm and lastly Misc classifier using Hyper Pipes algorithm.

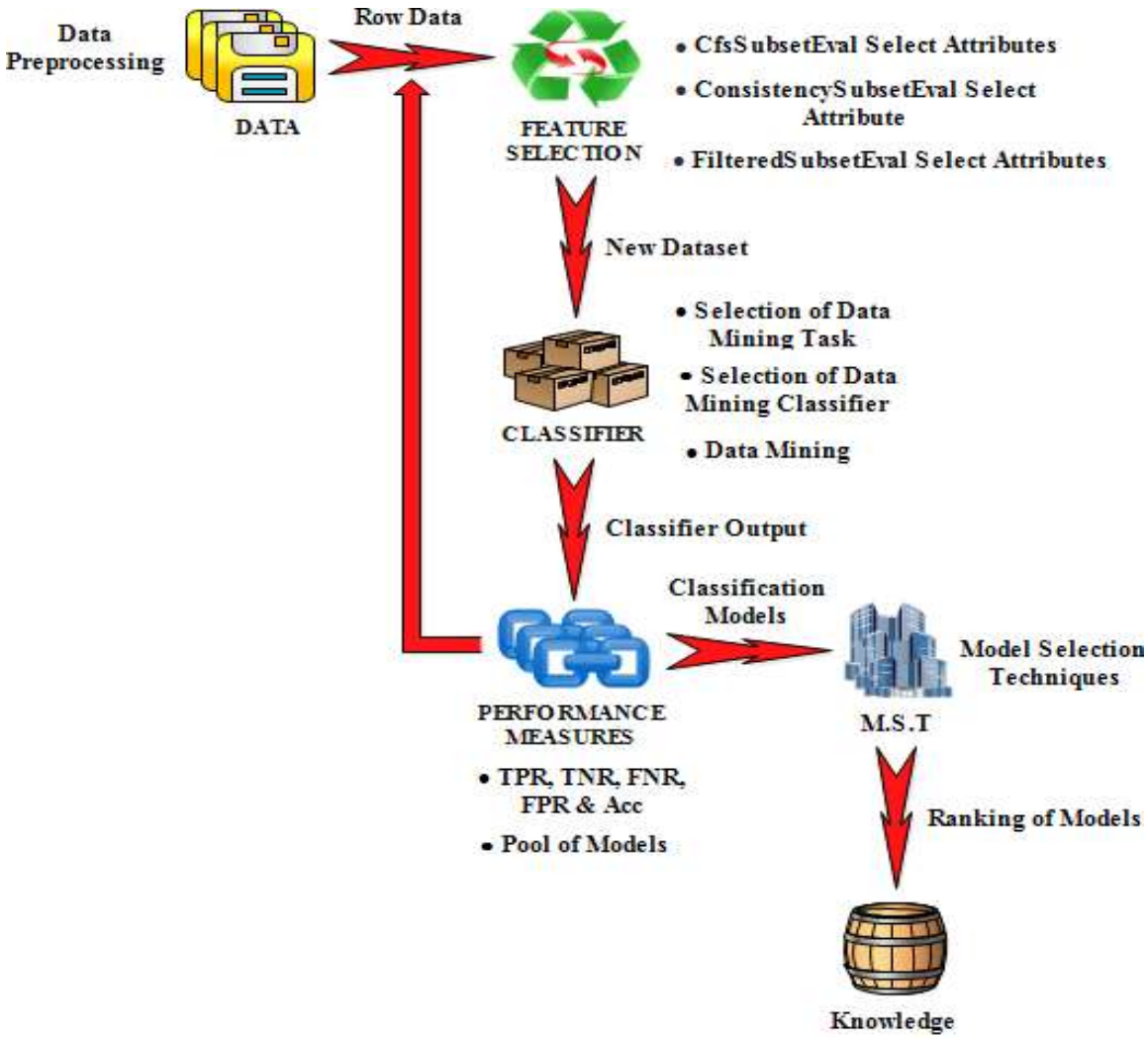

Fig.1. Research framework

The dataset of the research is from 2002 to 2013, it is a monthly data record merged into a single file. It has 301 instances (rows) and 54 attributes (columns). The missing values of the data were addressed by using the XLStart tool. After the data was normalized and the missing values were found, the data were stored in the database.

Fig. 1 shows the framework of the research which begins with an input where the data is preprocessed, and then passed to the classifier. After the data was classified by the classifier, the results will be stored in the classifier output. The performance measures will be concentrated and focused on TPR, TNR and Acc. The model selection techniques (MST) will use the weights and the performance measures to compute the result. The model selection techniques (MST) will finally rank the pools of model according to those who obtained the highest number of TPR, TNR and Acc.

2.1. Performance Measure

Confusion matrix is the raw yield (output) produced from a classification model. From the

confusion matrix, many performance measures can be computed such as true positive rate (TPR), true negative rate (TNR), false negative rate (FNR), false positive rate (FPR) and Accuracy [8]. The columns of the matrix are the actual classes and the rows are the predicted classes, the classifier output is evaluated by a confusion matrix as shown in Table 1. From Table 1, Equation (1)-(5) can be derived.

True Positive Rate is the percentage of right forecast for the positive class, e.g. Yes. False Positive Rate is the percentage of wrong forecast for the negative class, e.g. No. False Negative Rate is the percentage of wrong forecast for the positive class. True Negative Rate is the percentage of right forecast for the negative class. Accuracy is the percentage of right forecast for all classes. A standout amongst the most fundamental is the decision of a chosen measure so as to appropriately assess the classification performance and order the algorithms. In learning particularly imbalanced information, the general classification accuracy is frequently not a suitable measure of performance. At present, a paltry classifier that anticipates each case as the greater part class can attain high exactness [9-12].

Table 1. Confusion matrix

	Actual Positive Class	**Actual Negative Class**
Predicted Positive Class	TP	FP
Predicted Negative Class	FN	TN

$$TPR = TP / (TP + FN) \quad (1)$$

$$TNR = TN / (TN + FP) \quad (2)$$

$$FNR = FN / (FN + TP) \quad (3)$$

$$FPR = FP / (FP + TN) \quad (4)$$

$$Acc = (TP + TN) / (TP + FP + FN + TN) \quad (5)$$

2.2. Propose Model Selection Technique (MST)

The model selection technique is a method used to rank the pool of models and to increase the accuracy for imbalanced dataset in the classification model. MST was proposed based on weighted sum method by combining the performance measures for each class. In this paper, Acc, TPR and TNR were combined in order to get the highest TP and TN at the same time to maintain the highest Acc. The MST technique is as follows:

$$MST_m = \sum_{n=1} W_n P_n, \text{ for } m = 1,2,3,\ldots..x \quad (6)$$

where MST is the model selection technique that stands for ranking models, W_n is the relative weight for each performance measure and P_n is the results of each performance measures. The sum of the weight must to be 1 ($W_1 + W_2 + W_3$).

Let's take an example for us to understand the concept well. Let say 100 samples of blood

have been taken for an experiment to see whether they are diabetic or not. The samples are examined in a hospital by twodifferent doctors. The result for each doctor is given below.

Table 2. Result for Doctor A ($Model_1$)

	Actual Positive Class	**Actual Negative Class**
Predicted Positive Class	TP = 1	FP = 0
Predicted Negative Class	FN	TN

Table 3. Result for Doctor B ($Model_2$)

	Actual Positive Class	**Actual Negative Class**
Predicted Positive Class	TP = 10	FP = 10
Predicted Negative Class	FN = 0	TN = 80

From the classification results, Doctor A classifies that TP = 1, which means that 1 patient out of 10 is diabetic and the doctor said he is diabetic. For FN = 9, this means that 9 patients who are diabetic butthe doctor said they are not diabetic. Additionally, TN = 90 means 90 patient who are not diabetic and the doctor said they are not diabetic. Likewise, FP = 0 means 0 patient who is diabetic but the doctor said she is not diabetic.

Meanwhile, Doctor B classifies that TP = 10 means 10 patients who are diabetic and the doctor said they are diabetic (see Table 3). FN = 0 means there is no patient which is not diabetic and the doctor's experiment shows that she is not diabetic. Furthermore, TN = 80 means 80 patients who are not diabetic and the doctor said they are not diabetic. However, FP = 10 means 10 patients who are not diabetic but the doctor said they are diabetic. We can calculate the accuracy as follows:

Accuracy for $Model_1$ = (TP + TN) / (TP + TN + FN + FP) = (1 + 90) / (1 + 90 + 9 + 0) = 0.91

Accuracy for $Model_2$ = (TP + TN) / (TP + TN + FN + FP) = (10 + 80) / (10 + 80 + 0 +10) = 0.90

$Model_1$ obtained an accuracy of 0.91 and $Model_2$ obtained an accuracy of 0.90. By merely looking at these accuracies, $Model_1$ has the highest value compared to $Model_2$, therefore $model_1$ is the best. But, the truth is $Model_2$ is much better because in this case, the TPR is most important because it tell us that the doctor classifies well to determine the diabetic patients. The TPR for $Model_1$ is 10% but the TPR for $Model_2$ is 100%, which is far better. In this case, the most relevant model will be $Model_2$ although its Acc is slightly lower than $Model_1$. For $model_1$ the TPR and TNR can be calculated as:

TPR = TP / (TP + FN) = 1 / (1 + 9) = 0.10

TNR = TN / (TN + FP) = 90 / (90 + 0) = 1.00

For $model_2$ the TPR and TNR can be calculated as:

TPR = TP / (TP + FN) = 10 / (10 + 0) = 1.00

TNR = TN / (TN + FP) = 80 / (80 + 10) = 0.88

To find the most relevant model in a repository of models, MST proposed can be used. Based on Equation (6), MST would be calculated using these weights. (W_1 = 0.20, W_2 = 0.50 and W_3 = 0.30) The weights are assigned to focus on TPR followed by TNR and Acc.

MSTof $model_1$ = (W_1 * P_1) + (W_2 * P_2) + (W_3 * P_3)

= (0.2 * Acc) + (0.5 * TPR) + (0.3 * TNR) = (0.2 * 0.91) + (0.5 * 0.1) + (0.3 * 1) = 0.53

MST of $model_2$= (W_1 * P_1) + (W_2 * P_2) + (W_3 * P_3)

= (0.2 * Acc) + (0.5 * TPR) + (0.3 * TNR) = (0.2 * 0.90) + (0.5 * 1) + (0.3 * 0.88) = 0.94

Table 4 shows $Model_1$ is better than $Model_2$because it obtained the highest value of accuracy. However, now we can see that MST of $Model_1$ obtained 0.53 and MSTof $Model_2$ obtained 0.94. In summary, $Model_2$ is the optimal model not $Model_1$. This shows the disadvantage of relaying only on accuracy in the classification model. Also, it demonstrates that Doctor B's experiment is better than Doctor A's experiment.

Table 4. Confusion matrix

Model	Acc	TPR	TNR	MST	MST Ranking
M1	0.91	0.10	1.00	0.53	II
M2	0.90	1.00	0.88	0.94	I

3. RESULTS AND DISCUSSION

For this section, all the experiments will focus on three performance measures (Acc, TPR nd TNR) so as to test and see their consistency in the research. This paper will take 3 diverse weights in each and every model experiment, so as to test and see the effect of the weight diversity in the models. Each weight is assigned to the three performance measures (Acc, TPR and TNR). MST_1 focuses on Acc followed by TPR and TNR. Also, MST_2 focuses on TPR followed by Acc and TNR. Furthermore, MST_3 focuses on TNR followed by TPR and Acc. Our focus in this paper is MST_2 in order to achieve the highest TPR, as well as maintain its accuracy as discussed earlier. The weights are tabulated in Table 5.

Table 5. Assigned weight

	W_1 forAcc	W_1 forTPR	W_1 forTNR
MST_1	0.40	0.35	0.25
MST_2	0.30	0.50	0.20
MST_3	0.24	0.31	0.45

We generated 35 classification models with different classifiers and feature selection algorithms [4]. Experiments taken using ConsistencySubsetEval, CfsSubsetEval, FilteredSubsetEval as the feature selection algorithms. Table 6-9 show the results from the experiment conducted with different feature selection algorithms and using 7 classifiers as discussed in the methodology section.

We can see that Table 6 shows the summary of the experiments using 53 attributes. From the table, MST_3 has highest MST. At the same time, the model also has highest MST_1 and MST_3. Table 7 displays the brief of the experiments taken using ConsistencySubsetEval, Table 8using CfsSubsetEval and Table 9 using FilteredSubsetEval as the feature selection.

Table 6. Summary of the experiment 1 result using 53 attribute

Model	Acc	TPR	TNR	MST_1	MST_2	MST_3	MST Ranking
M_1	0.87	0.82	0.89	0.8567	0.8494	0.8622	II
M_2	0.87	0.78	0.91	0.8486	0.8331	0.8601	III
M_3	0.95	0.90	0.98	0.9408	0.9318	0.9475	I
M_4	0.84	0.81	0.86	0.8364	0.8312	0.8403	V
M_5	0.83	0.87	0.81	0.8371	0.8441	0.8319	IV
M_6	0.70	0.00	1.00	0.5291	0.4093	0.6174	VII
M_7	0.79	0.56	0.89	0.7351	0.6955	0.7642	VI

Table 7. Summary of experiment 3 result using ConsistencySubsetEvalselect attributes (DO Sat, DO, BOD, COD, SS, NH_3-NL, TEMP ^{O}C, COND, TUR, TS, NO_3, Ca and WQI class)

Model	Acc	TPR	TNR	MST_1	MST_2	MST_3	MST Ranking
M_{15}	0.87	0.77	0.92	0.8485	0.8306	0.8617	IV
M_{16}	0.89	0.82	0.91	0.8719	0.8611	0.8798	II
M_{17}	0.96	0.93	0.97	0.9538	0.9494	0.9571	I
M_{18}	0.84	0.80	0.86	0.8338	0.8266	0.8391	V
M_{19}	0.85	0.87	0.84	0.8548	0.8578	0.8525	III
M_{20}	0.70	0.00	1.00	0.5291	0.4093	0.6174	VII
M_{21}	0.76	0.93	0.69	0.8052	0.8344	0.7837	VI

Table 8. Summary of experiment 3 result using ConsistencySubsetEvalselect attributes (DO Sat, DO, BOD, COD, SS, NH_3-NL, TEMP ^{O}C, COND, TUR, TS, NO_3, Ca and WQI class)

Model	Acc	TPR	TNR	MST_1	MST_2	MST_3	MST Ranking
M_{22}	0.89	0.82	0.91	0.8719	0.8611	0.8798	IV
M_{23}	0.89	0.84	0.92	0.8795	0.8695	0.8870	II
M_{24}	0.97	0.95	0.99	0.9666	0.9617	0.9702	I
M_{25}	0.88	0.86	0.89	0.8748	0.8708	0.8777	III
M_{26}	0.85	0.87	0.84	0.8522	0.8558	0.8496	V
M_{27}	0.70	0.00	1.00	0.5291	0.4093	0.6174	VII
M_{28}	0.80	0.97	0.73	0.8434	0.8714	0.8227	VI

Table 9. Summary of experiment 5 result using FilteredSubsetEvalselect attributes (latitude, BOD, COD, SS, NH_3-NL, COND, TUR, As, Ca, Mg and WQI class)

Model	Acc	TPR	TNR	MST_1	MST_2	MST_3	MST Ranking
M_{29}	0.79	0.64	0.86	0.7536	0.7273	0.7731	III
M_{30}	0.78	0.64	0.84	0.7461	0.7215	0.7642	V
M_{31}	0.87	0.73	0.94	0.8379	0.8124	0.8567	I
M_{32}	0.83	0.77	0.85	0.8132	0.8033	0.8206	II
M_{33}	0.75	0.74	0.76	0.7473	0.7448	0.7492	IV
M_{34}	0.70	0.00	1.00	0.5291	0.4093	0.6174	VII
M_{35}	0.66	0.79	0.61	0.6951	0.7168	0.6790	VI

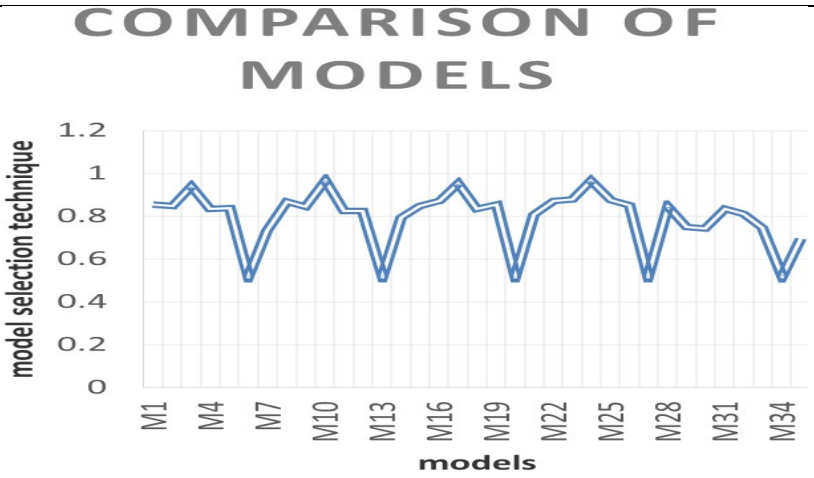

Fig.2. Graph of comparison of 35 models

Fig. 2 comes up with the comparison of all the 35 models of 5 different experiments using various feature selections. The graph shows that M_{10} is the optimal model since it obtained the highest MST_2 among the 35 models. The results in Table 6-9 show that M_{10} based on Simple

Logistic classifier is the best for MST_1, $MST2_2$ and MST_3. It can be concluded that M_3, M_{10}, M_7, M_{24} and M_{31} are the ones that have the highest TPR, TNR and Acc. In summary, it shows that the M_{10} is the optimal model among the 35 models which obtained the TPR = 0.96, TNR = 0.98 and Acc = 97.02%. The most significant features of the research are dissolve oxygen (DO), biochemical oxygen dissolve (BOD), chemical oxygen dissolve (COD), suspended solid (SS), pH value and ammoniacal nitrogen (NH_3-N). This model is proposed to be applied for classifying the water quality class of Kinta River, Perak, Malaysia.

4.CONCLUSION

The monthly data record of Kinta River is used from 2002-2013 to find the optimal model in the repository of models so as to determine the class of the River. The proposed model selection technique (MST) is used to rank the models. Results from the experiments, show that MST3 that focused on achieving highest in TNR (0.45), TPR (0.31) and Acc (0.24) able to find the relevant model form the repository. It is illustrated that the Function classifier using Simple logistic algorithm is the optimal model, having 97.02% accuracy. Future research work should focus on the effect of false negative rate (FNR) in classification models

5. ACKNOWLEDGEMENTS

This work has partly been supported by UniSZA (Grant No. RAGS /1/2014/ ICT07/ UniSZA/1/ RR095). The researchers are grateful toward the East Coast Environmental Research Institute (ESERI) UniSZA for providing the dataset.

6. REFERENCES

[1] Labatut V, Cherifi H. Evaluation of performance measures for classifiers comparison.Ubiquitous Computing and Communication Journal, 2011, 6:21-34

[2] Bennett PN, Dumais ST, Horvitz E. Probabilistic combination of text classifiers using reliability indicators: Models and results. In25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2002, pp. 207-214

[3] Ali S, Smith KA. On learning algorithm selection for classification. Applied Soft Computing, 2006, 6(2):119-138

[4] Muhammad SY, Makhtar M, Rozaimee A, Aziz AA, Jamal AA. Classification model for water quality using machine learning techniques. International Journal of Software Engineering and Its Applications, 2015, 9(6):45-52

[5] Lv Z, Luo S, Liu Y, Zheng Y. A new geometric approach to the complexity of model selection. In5th IEEE International Conference on Cognitive Informatics, 2006, pp. 268-273

[6] Ruta D, Gabrys B. Classifier selection for majority voting. Information Fusion, 2005, 6(1):63-81

[7] Chawla NV. Data mining for imbalanced datasets: An overview. In O. Maimon, & L. Rokach (Eds.), Data mining and knowledge discovery handbook. Massachusetts: Springer, 2009, pp. 875-886

[8] Makhtar M. Contributions of ensembles of models for predictive toxicology applications. PhD thesis, England: University of Bradford, 2012

[9] Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. Technical report, Berkeley: University of California, 2004

[10]Zhao H, Chen X, Nguyen T, Huang JZ, Williams G, Chen H. Stratified over-sampling bagging method for random forests on imbalanced data. InM. Chau, G. Wang, & H. Chen (Eds.), Pacific-Asia workshop on intelligence and security informatics. Cham: Springer, 2016, pp. 63-72

[11]Muchlinski D, Siroky D, He J, Kocher M. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. Political Analysis, 2015, 24(1):87-103

[12]Dittman DJ, Khoshgoftaar TM, Napolitano A. Is data sampling required when using random forest for classification on imbalanced bioinformatics data? In T. Bouabana-Tebibel, & S. Rubin (Eds.), Theoretical information reuse and integration. Cham: Springer, 2016, pp. 157-171

How to cite this article:
Makhtar M, Muhammad S Y, Rozaimee A, Shamsuddin S N W, Zamri N, Kamarudin M K A. Selection of classification models from repository of model for water quality dataset. J. Fundam. Appl. Sci., 2017, *9(6S), 751-761.*