A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS

Hossin, M.[1] and Sulaiman, M.N.[2]

[1]Faculty of Computer Science & Information Technology, Universiti Malaysia Sarawak,
94300 Kota Samarahan, Sarawak, Malaysia
[2]Faculty of Computer Science & Information Technology, Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia

ABSTRACT

Evaluation metric plays a critical role in achieving the optimal classifier during the classification training. Thus, a selection of suitable evaluation metric is an important key for discriminating and obtaining the optimal classifier. This paper systematically reviewed the related evaluation metrics that are specifically designed as a discriminator for optimizing generative classifier. Generally, many generative classifiers employ accuracy as a measure to discriminate the optimal solution during the classification training. However, the accuracy has several weaknesses which are less distinctiveness, less discriminability, less informativeness and bias to majority class data. This paper also briefly discusses other metrics that are specifically designed for discriminating the optimal solution. The shortcomings of these alternative metrics are also discussed. Finally, this paper suggests five important aspects that must be taken into consideration in constructing a new discriminator metric.

KEYWORDS

Evaluation Metric, Accuracy, Optimized Classifier, Data Classification Evaluation

1. INTRODUCTION

In data classification problems, data can be divided into commercial data, texts, DNAs and images. This paper emphasizes on commercial data as the focus of discussion. Furthermore, data classification can be divided into binary, multiclass and multi-labelled classification [33]. In this paper, the study is aimed on binary and multiclass classification which focuses on the evaluation metrics for evaluating the effectiveness of classifiers. In general, the evaluation metric can be described as the measurement tool that measures the performance of classifier. Different metrics evaluate different characteristics of the classifier induced by the classification algorithm.

From the literature, the evaluation metric can be categorized into three types, which are threshold, probability and ranking metric [2]. Each of these types of metrics evaluates the classifier with different aims. Furthermore, all these types of metrics are scalar group method where the entire performance is presented using a single score value. Thus, it makes easier to do the comparison and analysis although it could mask subtle details of their behaviours. In practice, the threshold and ranking metric were the most common metrics used by researchers to measure the performance of classifiers. In most cases, these types of metrics can be employed into three different evaluation applications [23].

First, the evaluation metrics were used to evaluate the generalization ability of the trained classifier. In this case, the evaluation metric is used to measure and summarize the quality of trained classifier when tested with the unseen data. Accuracy or error rate is one of the most

DOI : 10.5121/ijdkp.2015.5201

common metrics in practice used by many researchers to evaluate the generalization ability of classifiers. Through accuracy, the trained classifier is measured based on total correctness which refers to the total of instances that are correctly predicted by the trained classifier when tested with the unseen data.

Second, the evaluation metrics were employed as an evaluator for model selection. In this case, the evaluation metric task is to determine the best classifier among different types of trained classifiers which focus on the best future performance (optimal model) when tested with unseen data. Third, the evaluation metrics were employed as a discriminator to discriminate and select the optimal solution (best solution) among all generated solutions during the classification training. For example, the accuracy metric is employed to discriminate every single solution and select the best solution that produced by a particular classification algorithm. Only the best solution which is believed the optimal model will be tested with the unseen data.

For the first and second application of evaluation metrics, almost all types of threshold, probability and ranking metrics could be applied to evaluate the performance and effectiveness of classifiers. Conversely, only few types of metrics could be employed as a discriminator to discriminate and select the optimal solution during the classification training. This paper emphasizes on the third application of evaluation metrics for Prototype Selection (PS) classifiers. Only relevant evaluation metrics which are associated with the latter application of metrics are discussed.

In general, PS classifier is a generative type of classification algorithms that aim to generate a classifier model by applying sampling technique and simultaneously used the generated model to achieve the highest possible classification accuracy when dealing with the unseen data. Most of PS classifiers such as Monte Carlo Sampling (MCS) algorithm [32] genetic algorithm [22], evolutionary algorithm [10], and tabu search [38] were developed based on statistics, nature-inspired, optimization methods or combination of these methods. Basically, all of these classification algorithms applied stochastic or heuristic search to locate the optimal solution (a set of prototypes) by transforming the searching and discriminating processes into optimization problem. In other words, these algorithms begin with constructing and searching a fixed number of prototypes (candidate solution). Then, every produced solution will be evaluated in order to determine the optimal solution that best represents the training data and simultaneously aim to achieve better generalization ability when dealing with the unseen data. In order to search and discriminate the optimal solution from the large space of solutions, the selection of proper metric is crucial in discriminating a bulk of generated solutions. Without a proper and suitable evaluation metric, a particular PS classifier may obtain poor generalization ability when tested with the unseen data.

Typically, most of the PS classifiers employ the accuracy or the error rate (1-accuracy) to discriminate and to select the best (optimal) solution. However, using the accuracy as a benchmark measurement has a number of limitations. In [30, 37], they have demonstrated that the simplicity of this accuracy could lead to the suboptimal solutions especially when dealing with imbalanced class distribution. Furthermore, the accuracy also exhibits poor discriminating values to discriminate better solution in order to build an optimized classifier [17].

The purpose of this paper is to review and analyse all related evaluation metrics that were specifically designed for optimizing the PS classifiers. This paper begins with thorough reviews on commonly threshold type metrics and other metrics that are specifically used as a discriminator for discriminating the optimal solution for PS classifier. This section also discusses the limitations of these metrics as a discriminator in discriminating the optimal solution. This paper also recommends several important aspects in constructing a new discriminator metric for PS classifier. Finally, this paper ends with conclusions.

2. REVIEW OF DISCRIMINATOR METRICS

In a typical data classification problem, the evaluation metric has been employed into two stages, which are training stage (learning process) and testing stage. In training stage, the evaluation metric was used to optimize the classification algorithm. In other words, the evaluation metric was employed as the discriminator to discriminate and to select the optimal solution which can produce a more accurate prediction of future evaluation of a particular classifier. Meanwhile, in the testing stage, the evaluation metric was used as the evaluator to measure the effectiveness of produced classifier when tested with the unseen data.

As mentioned earlier, the interest of this paper is to review the use of evaluation metrics in discriminating and selecting the optimal solution in order to build optimized PS classifiers. In previous studies, there were various types of evaluation metrics that can be used to evaluate the quality of classifiers with different aims. On the contrary, there were little efforts have been dedicated to study and construct the metrics that are specifically designed to discriminate the optimal solution during the data classification training especially for Prototype Selection classifiers. On top of that, most of the previous studies were focused on binary classification problems as their main study domain [8]. Therefore, due to limited resources, this paper attempts to give the best reviews on the related studies as discussed below.

2.1. Threshold Types of Discriminator Metrics

For binary classification problems, the discrimination evaluation of the best (optimal) solution during the classification training can be defined based on confusion matrix as shown in Table 1. The row of the table represents the predicted class, while the column represents the actual class. From this confusion matrix, *tp* and *tn* denote the number of positive and negative instances that are correctly classified. Meanwhile, *fp* and *fn* denote the number of misclassified negative and positive instances, respectively. From Table 1, several commonly used metrics can be generated as shown in Table 2 to evaluate the performance of classifier with different focuses of evaluations. Due to multiclass problems, few of metrics listed in Table 2 have been extended for multi-class classification evaluations (see the last four metrics).

Table 1. Confusion Matrix for Binary Classification and the Corresponding Array Representation used in this Study

	Actual Positive Class	**Actual Negative Class**
Predicted Positive Class	True positive (*tp*)	False negative (*fn*)
Predicted Negative Class	False positive (*fp*)	True negative (*tn*)

As shown in the previous studies [3, 11, 16, 30], the accuracy is the most used evaluation metric in practice either for binary or multi-class classification problems. Through accuracy the quality of produced solution is evaluated based on percentage of correct predictions over total instances. The complement metric of accuracy is error rate which evaluates the produced solution by its percentage of incorrect predictions. Both of these metrics were used commonly by researchers in practice to discriminate and select the optimal solution.

The advantages of accuracy or error rate are, this metric is easy to compute with less complexity; applicable for multi-class and multi-label problems; easy-to-use scoring; and easy to understand by human. As pointed out by many studies, the accuracy metric has the limitations in evaluation and discrimination processes. One of the main limitations of accuracy is it produces less distinctive and less discriminable values [18, 29]. Consequently, it leads to less discriminating power to accuracy in selecting and determining the optimal classifier. In addition, the accuracy also powerless in terms of informativeness [25, 36] and less favour towards minority class instances [3, 9, 16, 30, 37].

Table 2. Threshold Metrics for Classification Evaluations

Metrics	Formula	Evaluation Focus
Accuracy (acc)	$\frac{tp+tn}{tp+fp+tn+fn}$	In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.
Error Rate (err)	$\frac{fp+fn}{tp+fp+tn+fn}$	Misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated.
Sensitivity (sn)	$\frac{tp}{tp+fn}$	This metric is used to measure the fraction of positive patterns that are correctly classified
Specificity (sp)	$\frac{tn}{tn+fp}$	This metric is used to measure the fraction of negative patterns that are correctly classified.
Precision (p)	$\frac{tp}{tp+fp}$	Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.
Recall (r)	$\frac{tp}{tp+tn}$	Recall is used to measure the fraction of positive patterns that are correctly classified
F-Measure (FM)	$\frac{2*p*r}{p+r}$	This metric represents the harmonic mean between recall and precision values
Geometric-mean (GM)	$\sqrt{tp*tn}$	This metric is used to maximize the *tp* rate and *tn* rate, and simultaneously keeping both rates relatively balanced
Averaged Accuracy	$\frac{\sum_{i=1}^{l}\frac{tp_i+tn_i}{tp_i+fn_i+fp_i+}}{l}$	The average effectiveness of all classes
Averaged Error Rate	$\frac{\sum_{i=1}^{l}\frac{fp_i+fn_i}{tp_i+fn_i+fp_i+}}{l}$	The average error rate of all classes
Averaged Precision	$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fp_i}}{l}$	The average of per-class precision
Averaged Recall	$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fn_i}}{l}$	The average of per-class recall
Averaged F-Measure	$\frac{2*p_M*r_M}{p_M+r_M}$	The average of per-class F-measure

Note: - each class of data; tp_i - true positive for C_i; fp_i - false positive for C_i; fn_i – false negative for C_i; tn_i - true negative for C_i; and M macro-averaging.

Instead of accuracy, the FM and GM also reported as a good discriminator and performed better than accuracy in optimizing classifier for binary classification problems [20]. To the best of our knowledge, no previous work has employed the FM and GM to discriminate and select the optimal solution for multiclass classification problems.

In contrast, the rest of metrics in Table 2 are unsuitable to discriminate and select the optimal solution due to single evaluation task (either positive or negative class). For discriminating and selecting the optimal solution during the classification training, the significance trade-off between classes is essential to ensure every class is represented by its representative prototype(s). The trade-off between classes becomes more crucial when imbalanced class data were used. The best selected solution turn out to be useless if none of the minority class instances were able to correctly predicted by the chosen representative(s) (prototype) or selected as the representative(s) (i.e. if using randomly representative (prototype) selection method) during the classification training.

2.2. Mean Square Error (MSE)

Supervised Learning Vector Quantization (LVQ) [21] is one of the Prototype Selection classifiers. During the learning process, supervised LVQ uses MSE to evaluate its performances during the classification training. In general, the MSE measures the difference between the predicted solutions and desired solutions. The smaller MSE value is required in order to obtain a better trained of supervised LVQ. The MSE is defined as below:

$$MSE = \frac{1}{n}\sum_{j=1}^{n}\left(P_j - A_j\right)^2 \tag{1}$$

where P_j is the predicted value of instance j, A_j is real target value of instance j and n is the total number of instances. Through the learning process of LVQ, the solution that has minimum MSE score will be used as the final model (best solution).

Similar to accuracy, the main limitation of MSE is this metric does not provide the trade-off information between class data. This may lead the discrimination process to select the sub-optimal solution. Moreover, this metric is really dependent on the weight initialization process. In extremely imbalanced class problem, if the initial weights are not proper selected (i.e. no initial weight to represent the minority class data), this may lead the discrimination process ends up with sub-optimal solution due to lack information of minority class data although the MSE value is minimized (under-fitting or over-fitting).

2.3. Area under the ROC Curve (AUC)

AUC is one of the popular ranking type metrics. In [13, 17, 31] the AUC was used to construct an optimized learning model and also for comparing learning algorithms [28,29]. Unlike the threshold and probability metrics, the AUC value reflects the overall ranking performance of a classifier. For two-class problem [13], the AUC value can be calculated as below

$$AUC = \frac{S_p - n_p(n_n + 1)/2}{n_p n_n} \tag{2}$$

where, S_p is the sum of the all positive examples ranked, while n_p and n_n denote the number of positive and negative examples respectively. The AUC was proven theoretically and empirically better than the accuracy metric [17] for evaluating the classifier performance and discriminating an optimal solution during the classification training.

Although the performance of AUC was excellent for evaluation and discrimination processes, the computational cost of AUC is high especially for discriminating a volume of generated solutions of multiclass problems. To compute the AUC for multiclass problems the time complexity is $O(|C|n \log n)$ for Provost and Domingos AUC model [28] and $O(|C|^2\ n \log n)$ for Hand and Till AUC model [13].

2.4. Hybrid Discriminator Metrics

Optimized Precision is a type of hybrid threshold metrics and has been proposed as a discriminator for building an optimized heuristic classifier [30]. This metric is a combination of accuracy, sensitivity and specificity metrics. The sensitivity and specificity metric were used for stabilizing and optimizing the accuracy performance when dealing with imbalanced class of two-class problems. The OP metric can be defined as below

$$OP = acc - \frac{|sp - sn|}{sp + sn} \tag{3}$$

where *acc* is the accuracy score, while *sp* and *sn* denotes specificity and sensitivity score respectively. In [30], the OP metric was able to discriminate and select a better solution and increase the classification performance of ensemble learners and Multi-Classifier Systems for solving Human DNA Sequences dataset.

Optimized accuracy with recall and precision (OARP) is another type of hybrid threshold metrics that is specifically designed as a discriminator to train the Monte Carlo Sampling (MCS) classifier during the classification training. There are two types of OARP; the Optimized Accuracy with Extended Recall-Precision version 1 (OAERP1) [15] and Optimized Accuracy with Extended Recall-Precision version 2 (OAERP2) [14, 16].

In general, both hybrid metrics are a combination of accuracy with extended recall (*rc*) and extended precision (*pr*) metric. The difference between both metrics is their Relationship Index (RI). In OAERP1, the RI is calculated based on correlation from [34], while OAERP 2 adopted the correlation given by [24]. The OAERP1 is formulated as follows:

$$OAERP = acc - RI_1 \tag{4}$$

where

$$RI_1 = \frac{|ep_1 + ep_2| - |er_1 + er_2|}{|ep_1 + ep_2| + |er_1 + er_2|} \tag{5}$$

Meanwhile, the OAERP2 is formulated as follows:

$$OAERP = acc - RI_2 \tag{6}$$

where

$$RI_2 = \frac{\left|\frac{ep_1 - er_2}{ep_1 + er_2}\right| - \left|\frac{ep_2 - er_1}{ep_2 + ec_1}\right|}{2} \tag{7}$$

For both RIs formula, the *ep* and *er* represent extended precision and extended recall respectively and the numbering denotes class 1 (positive class) and class 2 (negative class).

As shown in [14, 15, 16], the RI value for both metrics have the possibility to return zero score. If this happens, the OAERP1 and OAERP2 score are presumed equivalent to accuracy score. Besides, the OAERP1 and OAERP2 score also can return a negative score. To avoid negative score, the RI score needs to be resized using decimal scaling method. Both metrics have demonstrated better performance than accuracy in terms of distinctiveness and discriminable of produced-value. However, the OEARP2 shows better discriminating power to choose an optimal solution and able to build a better trained of MCS classifier [14, 16].

The main limitation of these three hybrid metrics is its only limited for discriminating and evaluating the binary classification problems. In real-world dataset, the data available is not limited to two-class problem. Many datasets comprise more than two classes. To the best of our knowledge, no previous work has modified these metrics for evaluating the multiclass data. Thus, the effectiveness of these metrics is still questionable for multiclass classification problems.

2.5. Other Metrics

Instead of the abovementioned metrics, there are graphical-based metrics, which are better than accuracy, have been proposed to evaluate the performance of classifiers. As reported in [27], these metrics able to depict the trade-offs between different evaluation perspectives which allowing richer analysis of results. Although these metrics are better than accuracy or error rate, its graphical-based output limits a metric such as Receiver Operating Curve (ROC) [4], Bayesian Receiver Operating Characteristic (B-ROC) [1], Precision-Recall Curve [5], cost curve [6], lift and chart calibration plot [35] to be employed as a discriminator.

Besides, there are few other metrics that were specifically designed for a particular classification algorithm. Information gain and entropy metrics are two probability types of metrics that were used for evaluating the utility of attributes of data in building the optimized decision tree classifiers [26]. Due to specific purpose, these metrics are unsuitable for adoption to discriminate and select the optimal solution. To the best of our knowledge, no previous work has exploited these metrics to discriminate a bulk of generated solutions during the classification training of PS classifiers.

3. Important Factors in Constructing New Metrics

Through the reviews processed, this paper has figured out several factors that might help the researchers in designing and constructing a new metric or choosing the suitable metric for discriminating the optimal solution of PS classification algorithms. The lists of these important factors are briefly described as below.

1. Issue on multiclass problem

   Many of current metrics were originally developed and applicable for binary classification problems with different tasks of evaluation. This is the major limitation that restricted many good metrics for widely used as a discriminator in discriminating the optimal solution. In reality, the data available are not limited to two-class problem. Most of data involves more than two classes. For example, the student grades can be categorized into A, B, C, D, E and F. Therefore, the future development of new metric or choosing a suitable metric should accommodate this issue into consideration.

2. Less complexity and less computational cost

   Since data nowadays involve multiclass data the used of particular metric becomes more complex due to increasing classes that need to be evaluated. As the consequences, it produces high computational cost and affects the classification training speed. Due to this matter, most researchers simply applied accuracy or error rate to discriminate their produced solutions. Although, accuracy and error rate is less complex metric and easy-to-use score, the PS classifiers still require longer learning process or training process. Therefore, the biggest challenge of future development of a new metric is to design and construct a less complex metric with less computational cost and comprehensible enough to discriminate an optimal solution from a bulk of generated solutions.

3. Distinctiveness and discriminable

Less distinctiveness and discriminable value of produced solution is another drawback of accuracy metric [15]. As a result, this drawback will cause the discrimination and searching process of an optimal solution easily trapped at local optima (plateau). Thus, this drawback must be avoided by any discriminator. In other words, the development of future metrics must be able to produce a distinctive and discriminable value for better searching and discriminating the optimal solution in huge solution space. The details of this effect in discriminating the optimal solution are discussed in Table 3 and Table 5.

4. Informativeness

Another drawback of many current metrics is there is no trade-off information between classes [24, 35]. For example, the most popular metric accuracy could not discriminate the good and bad (informative and non-informative) solutions especially when two or more solutions are equivalent or even contradict as shown in Table 3 and 4 respectively.

From Table 3, the accuracy metric could not distinguish which solution is better due to non-distinctiveness and non-discriminable produced value. Intuitively, solution a_2 is better than a_1 since a_2 can predict correctly all minority class members. In a_1, there is none of the minority class member is correctly predicted, which conclude a_1 is a poor solution. On the other hands, in Table 4, the accuracy metric concludes solution a_1 is better than a_2 through score comparison. However, a_1 is a poor solution where none of minority class member is correctly predicted by a_1. Intuitively, solution a_2 shows better result although the score is lower than a_1. In this case, solution a_2 able to predict correctly all minority class members as compared to a_1.

From both examples, it shows that the informativeness aspect is essential feature for any metric in discriminating the informative and optimal solution.

Table 3. Informativeness Analysis for Binary Classification Problem using Imbalanced Class Distribution (5:95) with Two Equivalent Solutions

sol	*tp*	*fp*	*tn*	*fn*	*total*	*accuracy*
a_1	0	5	95	0	95	0.9500
a_2	5	0	90	5	95	0.9500

Table 4. Informativeness Analysis for Binary Classification Problem using Imbalanced Class Distribution (5:95) with Two Contradictory Solutions

sol	*tp*	*fp*	*tn*	*fn*	*total*	*accuracy*
a_1	0	0	95	5	95	0.9500
a_2	5	6	89	0	94	0.9400

5. Favour towards the minority class

According to [11, 12, 19], the most popular accuracy metric is greatly affected by the proportion of majority class and less impact on minority class. Hence, it is important to employ a proper evaluation metric that could favor towards the minority class than the majority class. In other words, the more minority class instances are correctly predicted, the better solution is produced especially for extremely imbalanced class problems. From demonstrated example in Table 5, any good metric or discriminator should rank the solution as follow: $a_6 \rightarrow a_5 \rightarrow a_4 \rightarrow a_3 \rightarrow a_2 \rightarrow a_1$. As demonstrated in Table 5, intuitively, the solution a_6

is the most informative solution and more favor towards minority class, while a_1 is the poorest solution since it has none of minority class members. Based on accuracy metric scores, none of these solutions could be ranked as suggested due to equivalent score among all solutions. Furthermore, Table 5 also shows that the accuracy metric produced less distinctive and less discriminable score which can cause the accuracy metric easily trapped at local optima during the searching of an optimal solution.

Table 5. Favors towards Minority Class Analysis for Binary Classification Problem using Extremely Imbalanced Class Distribution (5:9995)

sol	*tp*	*fp*	*tn*	*fn*	*total*
a_1	0	0	9995	5	9995
a_2	1	1	9994	4	9995
a_3	2	2	9993	3	9995
a_4	3	3	9992	2	9995
a_5	4	4	9991	1	9995
a_6	5	5	9990	0	9995

4. CONCLUSIONS

The selection of suitable metric for discriminating the optimal solution in order to obtain an optimized classifier is a crucial step. The proper selection of metric will ensure that the classification training of generative type classifier is optimal. This paper hopes that with the reviews of some metrics for discriminating the optimal solution will sensitize the data mining researchers to the issue and encourage the researchers to think carefully, prior to select and apply the suitable metric for optimizing the classification training. Besides, this paper also suggested several important aspects in constructing a better metric for discriminating the optimal solution for generative type of classification algorithms.

ACKNOWLEDGEMENTS

This work has been funded by the Ministry of Education, Malaysia under SLAI scholarship. We also thanks to Universiti Malaysia Sarawak and other reviewers for their supports and helpful comments.

REFERENCES

[1] A.A. Cardenas and J.S. Baras, “B-ROC curves for the assessment of classifiers over imbalanced data sets”, in Proc. of the 21st National Conference on Artificial Intelligence Vol. 2, 2006, pp. 1581-1584

[2] R. Caruana and A. Niculescu-Mizil, “Data mining in metric space: an empirical analysis of supervised learning performance criteria”, in Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04), New York, NY, USA, ACM 2004, pp. 69-78.

[3] N.V. Chawla, N. Japkowicz and A. Kolcz, “Editorial: Special issue on learning from imbalanced data sets”, SIGKDD Explorations, 6 (2004) 1-6.

[4] T. Fawcett, “An Introduction to ROC Analysis”, Pattern Recognition Letters, 27 (2006) 861-874.

[5] J. Davis and M. Goadrich, “The relationship between precision-recall and ROC curves”, in Proc. of the 23rd International Conference on Machine Learning, 2006, pp. 233-240.

[6] C. Drummond, R.C. Holte, “Cost curves: An Improved method for visualizing classifier performance”, Mach. Learn. 65 (2006) 95-130.

[7] P.A. Flach, P.A., “The Geometry of ROC Space: understanding Machine Learning Metrics through ROC Isometrics”, in T. Fawcett and N. Mishra (Eds.) Proc. of the 20th Int. Conference on Machine Learning (ICML 2003), Washington, DC, USA, AAAI Press, 2003, pp. 194-201.

[8] V. Garcia, R.A. Mollineda and J.S. Sanchez, “A bias correction function for classification performance assessment in two-class imbalanced problems”, Knowledge-Based Systems, 59(2014) 66-74.

[9] S. Garcia and F. Herrera, “Evolutionary training set selection to optimize C4.5 in imbalance problems”, in Proc. of 8th Int. Conference on Hybrid Intelligent Systems (HIS 2008), Washington, DC, USA, IEEE Computer Society, 2008, pp.567-572.

[10] N. Garcia-Pedrajas, J. A. Romero del Castillo and D. Ortiz-Boyer, “A cooperative coevolutionary algorithm for instance selection for instance-based learning”. Machine Learning (2010), 78 (2010) 381-420.

[11] Q. Gu, L. Zhu and Z. Cai, “Evaluation Measures of the Classification Performance of Imbalanced Datasets”, in Z. Cai et al. (Eds.) ISICA 2009, CCIS 51. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 461-471.

[12] S. Han, B. Yuan and W. Liu, “Rare Class Mining: Progress and Prospect”, in Proc. of Chinese Conference on Pattern Recognition (CCPR 2009), 2009, pp. 1-5

[13] D. J. Hand and R. J. Till, “A simple generalization of the area under the ROC curve to multiple class classification problems”, Machine Learning, 45 (2001) 171-186.

[14] M. Hossin, M. N. Sulaiman, A. Mustapha, and N. Mustapha, “A Novel Performance Metric for Building an Optimized Classifier”, Journal of Computer Science, 7(4) (2011) 582-509.

[15] M. Hossin, M. N. Sulaiman, A. Mustapha, N. Mustapha and R. W. Rahmat, “ OAERP: a Better Measure than Accuracy in Discriminating a Better Solution for Stochastic Classification Training”, Journal of Artificial Intelligence, 4(3) (2011) 187-196.

[16] M. Hossin, M. N. Sulaiman, A. Mustapha, N. Mustapha and R. W. Rahmat, “A Hybrid Evaluation Metric for Optimizing Classifier”, in Data Mining and Optimization (DMO), 2011 3rd Conference on, 2011, pp. 165-170.

[17] J. Huang and C. X. Ling, “Using AUC and accuracy in evaluating learning algorithms”, IEEE Transactions on Knowledge Data Engineering, 17 (2005) 299-310.

[18] J. Huang and C. X. Ling, “Constructing new and better evaluation measures for machine learning”, in R. Sangal, H. Mehta and R. K. Bagga (Eds.) Proc. of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp.859-864.

[19] N. Japkowicz, “Assessment metrics for imbalanced learning”, in Imbalanced Learning: Foundations, Algorithms, and Applications, Wiley IEEE Press, 2013, pp. 187-210

[20] M. V. Joshi, “On evaluating performance of classifiers for rare classes”, in Proceedings of the 2002 IEEE Int. Conference on Data Mining (ICDN 2002) ICDM’02, Washington, D. C., USA: IEEE Computer Society, 2002, pp. 641-644.

[21] T. Kohonen, Self-Organizing Maps, 3rd ed., Berlin Heidelberg: Springer-Verlag, 2001.

[22] L. I. Kuncheva and J. C. Bezdek, “Nearest Prototype Classification: Clustering, Genetic Algorithms, or Random Search?” IEEE Transactions on Systems, Man, and Cybernetics-Part C: Application and Reviews, 28(1) (1998) 160-164.

[23] N. Lavesson, and P. Davidsson, “Generic Methods for Multi-Criteria Evaluation”, in Proc. of the Siam Int. Conference on Data Mining, Atlanta, Georgia, USA: SIAM Press, 2008, pp. 541-546.

[24] P. Lingras, and C. J. Butz, “Precision and recall in rough support vector machines”, in Proc. of the 2007 IEEE Int. Conference on Granular Computing (GRC 2007), Washington, DC, USA: IEEE Computer Society, 2007, pp.654-654.

[25] D. J. C. MacKay, Information, Theory, Inference and Learning Algorithms. Cambridge, UK: Cambridge University Press, 2003.

[26] T. M. Mitchell, Machine Learning, USA: MacGraw-Hill, 1997.

[27] R. Prati, G. Batista, and M. Monard, “A survery on graphical methods for classification predictive performance evaluation”, IEEE Trans. Knowl. Data Eng. 23(2011) 1601-1618.

[28] F. Provost, and P. Domingos, “Tree induction for probability-based ranking”. Machine Learning, 52 (2003) 199-215.

[29] A. Rakotomamonyj, “Optimizing area under ROC with SVMs”, in J. Hernandez-Orallo, C. Ferri, N. Lachiche and P. A. Flach (Eds.) 1st Int. Workshop on ROC Analysis in Artificial Intelligence (ROCAI 2004), Valencia, Spain, 2004, pp. 71-80.

[30] R. Ranawana, and V. Palade, “Optimized precision-A new measure for classifier performance evaluation”, in Proc. of the IEEE World Congress on Evolutionary Computation (CEC 2006), 2006, pp. 2254-2261.

[31] S. Rosset, “Model selection via AUC”, in C. E. Brodley (Ed.) Proc. of the 21st Int. Conference on Machine Learning (ICML 2004), New York, NY, USA: ACM, 2004, pp. 89.

[32] D.B. Skalak, “Prototype and feature selection by sampling and random mutation hill climbing algorithm”, in W. W. Cohen and H. Hirsh (Eds.) Proc. of the 11th Int. Conference on Machine Learning (ICML 1994), New Brunswick, NJ, USA: Morgan Kaufmann, 1994, pp.293-301.

[33] M. Sokolova and G. Lapame, “A systematic analysis of performance measures for classification tasks”, Information Processing and Management, 45(2009) 427-437.

[34] P. N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Boston, USA: Pearson Addison Wesley, 2006.

[35] M. Vuk and T. Curk, “ROC curve, lift chart and calibration plot”, Metodoloˇski zvezki, 3(1) (2006) 89-108.

[36] H. Wallach, “Evaluation metrics for hard classifiers”. Technical Report. (Ed.: Wallach, 2006) http://www.inference.phy.cam.ac.uk/hmw26/papers

[37] S. W. Wilson, “Mining oblique data with XCS”, in P. L. Lanzi, W. Stolzmann and S. W. Wilson (Eds.) Advances in Learning Classifier Systems: Third Int. Workshop (IWLCS 2000), Berlin, Heidelberg: Springer-Verlag, 2001, pp. 283-290.

[38] H. Zhang and G. Sun, “Optimal reference subset selection for nearest neighbor classification by tabu search”, Pattern Recognition, 35(7) (2002) 1481-1490.

AUTHORS

Mohammad b. Hossin is a senior lecturer at Universiti Malaysia Sarawak. He received his B.IT (Hons) from Universiti Utara Malaysia (UUM) in 2000 and M.Sc. in Artificial Intelligence from University of Essex, UK in 2003. Then, in 2012, he received his Ph.D in Intelligent Computing from Universiti Putra Malaysia (UPM). His main research interests include data mining, decision support systems optimization using nature-inspired algorithms and e-learning.

Md Nasir Sulaiman received his Bachelor in Science with Education major in Mathematics from University Pertanian Malaysia in 1983. He received Master in Computing from University of Bradford, U.K., in 1986 and PhD degree in Computer Science from Loughborough University, U.K., in 1994. He is a lecturer at the University Putra Malaysia since 1986. He has been promoted to Associate Professor in 2002. His research interests are Intelligent Computing and Smart Home.