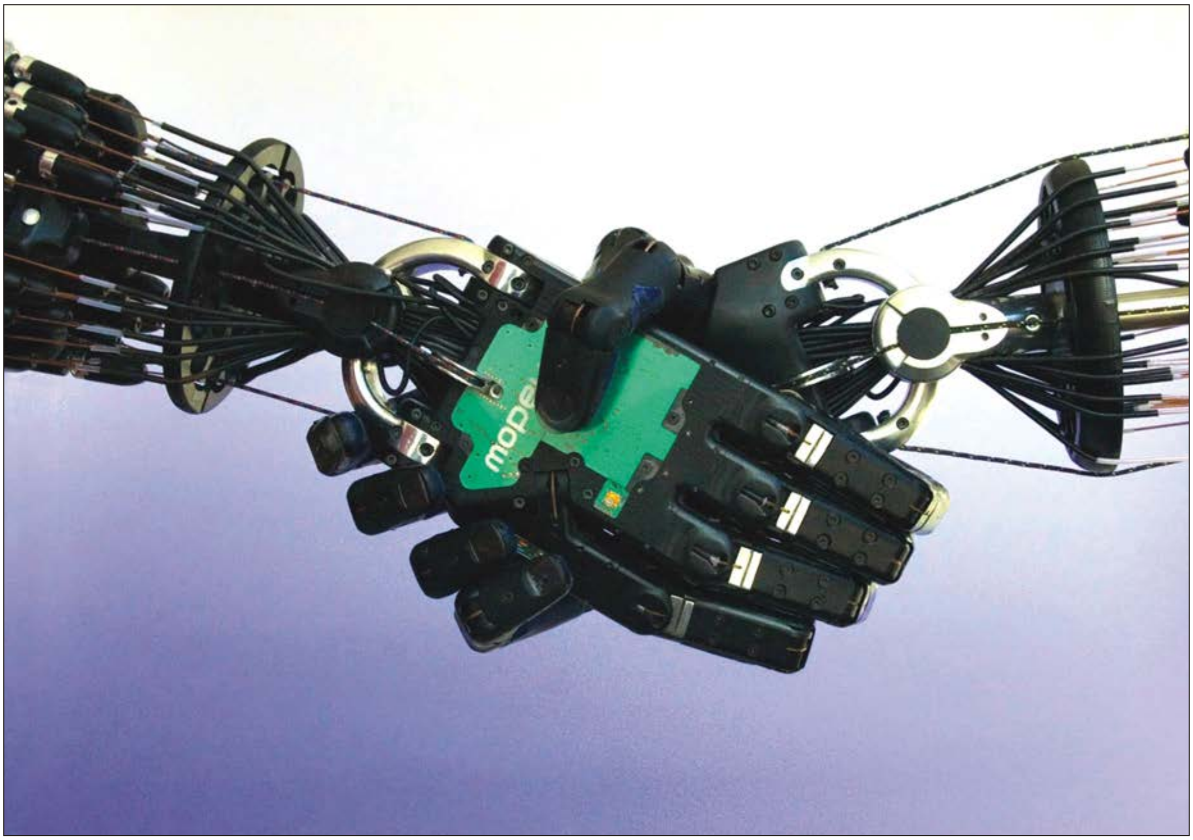

Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Artificial Intelligence 4 (3): 187-196, 2011
ISSN 1994-5450 / DOI: 10.3923/jai.2011.187.196

OAERP: A Better Measure than Accuracy in Discriminating a Better Solution for Stochastic Classification Training

[1,2]M. Hossin, [1]M.N. Sulaiman, [1]A. Mustapha, [1]N. Mustapha and [1]R.W. Rahmat
[1]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia
[2]Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, Malaysia

Corresponding Author: M. Hossin, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

ABSTRACT

The use of accuracy metric for stochastic classification training could lead the solution selecting towards the sub-optimal solution due to its less distinctive value and also unable to perform optimally when confronted with imbalanced class problem. In this study, a new evaluation metric that combines accuracy metric with the extended precision and recall metrics to negate these detrimental effects was proposed. This new evaluation metric is known as Optimized Accuracy with Extended Recall-precision (OAERP). By using two examples, the results has shown that the OAERP metric has produced more distinctive and discriminating values as compared to accuracy metric. This paper also empirically demonstrates that Monte Carlo Sampling (MCS) algorithm that is trained by OAERP metric was able to obtain better predictive results than the one trained by the accuracy metric alone, using nine medical data sets. In addition, the OAERP metric also performed effectively when dealing with imbalanced class problems. Moreover, the t-test analysis also shows a clear advantage of the MCS model trained by the OAERP metric against its previous metric over five out of nine medical data sets. From the abovementioned results, it is clearly indicates that the OAERP metric is more likely to choose a better solution during classification training and lead towards a better trained classification model.

Key words: Evaluation metric, hybrid evaluation metric, accuracy, recall, precision, stochastic classification model

INTRODUCTION

In the context of stochastic classification algorithms, a key objective of classification training is to identify and select the best solution among all generated solutions that well fits the input data (training data) and accurately predict the class labels of unknown data (test data). However, using accuracy metric as evaluator and discriminator to discriminate and select the best solution has limitations. Chawla *et al.* (2004), Garcia and Herrera (2008), Huang and Ling (2005), Ranawana and Palade (2006) and Wilson (2001) demonstrated the simplicity of this accuracy metric could lead to the sub-optimal solutions especially when dealing with imbalanced class problem. It is due to a minority class instances has very little impact on the accuracy as compared to the majority class instances. Furthermore, the accuracy metric also exhibits poor discriminating power to discriminate and select the best solution in order to build an optimized classification model due to its less distinctive value (Huang and Ling, 2007).

From the previous study, little efforts were dedicated to study and improve the discrimination process of classification training using an alternative evaluation metric. Ferri *et al*. (2002), Huang and Ling (2007) and Rakotomamonyj (2004) had proposed area under ROC (AUC) metric as the objective function for discriminating the best solution. The AUC was proven theoretically and empirically better not only to construct optimized learning models (Ferri *et al*., 2002; Huang and Ling, 2007; Rakotomamonyj, 2004; Yan *et al*., 2003) but also in evaluating classifiers (Bradley, 1997; Huang and Ling, 2005; Provost and Domingos, 2003). Although, the AUC performance is good, the computational cost of this metric is high in order to evaluate a volume of generated solutions. For instance to compute the AUC for multi-class problem the time complexity is $O(|C|n \log n)$ for Provost and Domingos AUC model (Provost and Domingos, 2000) and $O(|C|^2 n \log n)$ for Hand and Till AUC model (Hand and Till, 2001). In fact, for two-class problem a special algorithm is employed to calculate the AUC value (Fawcett, 2006).

From the literature, there was a study that has been conducted to improve the accuracy metric using hybridizing technique (Ranawana and Palade, 2006). The new hybridized metric was called Optimized Precision (OP) which a combination of accuracy metric with sensitivity and specificity metrics. In this study, 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 data set. To the best of our knowledge, there was no such efforts have been made to employ this evaluation metric to train other application domains or types of data set. As a result, the effectiveness of this evaluation metric is still uncertain and questionable for data classification.

Similar to OP metric (Ranawana and Palade, 2006), the main purpose of this study was to improve the problem of accuracy metric in discriminating the best solution in order to build an optimized stochastic classification models for data classification. This paper introduced a new hybridized performance metric derived from the combination of accuracy metric with the extended precision and extended recall metrics. The new performance metric is known as an optimized accuracy with extended recall-precision (OAERP) metric. In addition, the proposed metric is also expected to impose lighter computational complexity to facilitate the easy computation and adaptation in discriminating a volume of generated solution during the classification training.

MATERIALS AND METHODS

Background of research project: The research to be presented in this study is a part of doctoral research project conducted in Intelligent Computing Lab, Universiti Putra Malaysia (UPM). This project is entitled Optimizing Stochastic Classification Models via Novel Evaluation Metrics for Two-Class and Multi-Class Classification Problems and it was initiated on July 2008 and will be end on 2012.

Proposed evaluation metric: As aforesaid, the aim of this study was to propose a new evaluation metric which combined the accuracy metric with the extended version of precision and recall metrics. From the literature, precision and recall are two evaluation metrics that are commonly used as the alternative metrics to measure the performance of two-class classification problem for two different aspects (Buckland and Gey, 1994). Basically, precision is used to measure the fraction of positive data that are correctly predicted in a positive class (confidence) while recall measures the fraction of positive data being correctly classified over the total of positive data (coverage). From our point of view, the conventional precision and recall metrics are unsuitable for the combination process with accuracy metric. This is because both metrics only measure one class of data. According to Ranawana and Palade (2006), the ideal concept to construct a new evaluation metric is the proposed metric should be able to maximize every class data. Therefore, the extended precision and extended recall were proposed. Lingras and Butz (2007) extended the notion of this conventional precision and recall metrics by defining separate values of precision and recall for each class of data. Through these metrics, the performance of every class data could be measured individually and in the same time provides more information for evaluation purposes.

One might question, in what means the accuracy metric could be hybridized with the extended precision and extended recall metrics? To answer this question, two important formulas from Ranawana and Palade (2006), namely the Relationship Index (RI) and Optimized Precision (OP) were adopted. The details of these reference formulas can be found in Ranawana and Palade (2006). For combination process using both formulas, it involves two-step efforts. At the first step, an appropriate correlation between extended precision and extended recall need to be identified in order to apply the RI formula. Then, the next step is to identify the best means to adopt the OP formula in order to merge the accuracy metric with the RI formula.

According to Tan *et al.* (2006), the ideal concept for building an optimized classification model using the precision and recall metrics are by maximizing the both values of Precision (p) and recall (r) (p↑, r↑). Based on this correlation, the RI formula could be employed by dividing the difference of total precision value and total recall value with the summation of total precision value and total recall value. In short, RI can be defined as Eq. 1:

$$\mathrm{RI} = \frac{|(p_1 + p_2) - (r_1 + r_2)|}{(p_1 + p_2) + (r_1 + r_2)} \tag{1}$$

A low RI value entails a low $|(p_1 + p_2)\text{-}(r_1 + r_2)|$ and a high $(p_1 + p_2) + (r_1+r_2)$ values which indicate the values of each recall (r_1, r_2) and precision (p_1, p_2) in both classes are comparatively equivalent. The RI only returns value of zero whenever each precision and recall value in both classes is equal.

As mentioned earlier, the use of accuracy metric alone could lead the searching and discriminating (sub-solutions) process to be under-performing due to its less distinctive value. Due to this drawback, this leads us to combine the beneficial properties of RI with the accuracy metric as defined in Eq. 2. For simplicity, the new evaluation metric is called optimized accuracy with extended recall-precision (OAERP):

$$\mathrm{OAERP} = \mathrm{Acc\text{-}R1} \tag{2}$$

Unlike the accuracy value, we believed that the OAERP value is more distinctive and discriminative with the help of RI value. We also believed that the OAERP metric is able to perform effectively when dealing with any distribution class of data (balanced and imbalanced class distribution) since the performance of each class is incorporating to produce the OAERP value. In addition, by adopting the RI and OP formulas, the computation complexity of computing the OAERP value is still simple and moderate. In fact, all values required are simply derived from the confusion matrix and need not a special algorithm to compute the OAERP value.

Resizing and smoothing OAERP value: In calculating the OAERP value using Eq. 2, we observed that the OAERP value tends to result in negative and deviate too far from the actual accuracy value especially when the RI value is larger than or almost equivalent to the accuracy value. For that reason, we propose to resize RI value into a relatively small value before computing the OAERP value. To perform smoothing on the OAERP value, the decimal scaling method (Al-Shalabi, 2011) is used to resize the RI value as shown in Eq. 3:

$$RI_{new} = \frac{RI_{old}}{10^{x}} \tag{3}$$

where, x>0. In this study, we set the x = 1 for the entire experiment. By resizing the RI value, the resulting OAERP value will be comparatively closer to the accuracy value but still distinctive and discriminative. This claim will be validated in the experimental section.

Experimental methods: For comparison and discussion, this study has been limited by comparing the OAERP metric against the accuracy metric only. Moreover, the two-class classification problem was used for comparing both metrics. For the performance evaluation, two kinds of experiments were conducted to demonstrate the advantage of OAERP metric against the accuracy metric.

Experiment 1: For the first experiment, the OAERP metric was compared with the accuracy metric using two examples (case studies). Both metrics were analyzed manually in terms of distinctness of value produced and the performance ability when dealing with any data distribution (balanced and imbalanced). To ensure fair comparison, the intuition decision (common sense) was employed as baseline for comparison discussion (MacKay, 2003). This method is important in order to demonstrate that the evaluation made by the system is aligned with the human intuition in evaluating the best solution. In addition, we also restricted our discussion to the solutions that are indistinguishable based on the accuracy value. All data used in this comparison were manually generated and represented using confusion matrix as shown in Table 1.

Experiment 2: To demonstrate the applicability and advantage of OAERP metric over accuracy metric in discriminating the best solution, a naive instance selection algorithm which is Monte Carlo Sampling (MCS) algorithm (Skalak, 1994) was employed for the experimentation. MCS algorithm was chosen because this classification algorithm simply applies accuracy metric to discriminate the best solution during the classification training. However, as we observed, the best solution that discriminated and selected by accuracy metric does not always achieve better predictive result when tested with test data. In contrast, other generated solutions that obtained slightly lower training accuracy value than the best one able to produce better predictive result. On top of that, even if two solutions obtained equivalent training accuracy value, they may obtain different predictive results. For this reason, an appropriate evaluation metric which is more discriminating than accuracy metric should be applied to discriminate and select the best solution. Therefore, in this study, we propose this naive stochastic classification algorithm to be adopted for the second experiment.

Table 1: Confusion matrix

	Actual positive class	Actual negative class
Predicted positive class	True positive (TP)	False negative (FN)
Predicted negative class	False positive (FP)	True negative (TN)

Table 2: Brief description of each chosen data set

Dataset	NoI	NoA	MV	Minority class (%)	Majority class (%)
Breast cancer-Original (BCO)	699	9	Yes	34.48	65.52
Breast cancer-Diagnostic (BCD)	569	30	No	37.26	62.74
Breast cancer-Prognostic (BCP)	198	32	Yes	23.74	76.26
Heart	270	13	No	44.44	55.56
Hepatitis (Hepa)	155	19	Yes	20.65	79.35
Liver	345	6	No	42.20	57.80
Parkinson (Pksn)	197	22	No	25.38	74.62
Pima-Indian diabetes (Pima)	768	8	No	34.90	65.10
SPECTF-Heart (SPECTF)	267	44	No	35.58	64.42

NoI: No. of instances, NoA: No. of attributes, MV: Missing value

Data sets: For the purpose of comparison and evaluation on the performance of OAERP metric against the accuracy metric, nine medical data sets from (Frank and Asuncion, 2010) were selected. The data sets represent real-world problems and involve challenging issues such as imbalanced class problem. The chosen data sets also vary in terms of relative proportions between two classes and differ in terms of the number of attributes and instances. Brief descriptions about the selected data sets are summarized in Table 2. In pre-processing, all of the selected instances were normalized using min-max normalization method within the range of [0, 1] to prevent any attribute variables from dominating the analysis (Al-Shalabi, 2011). All missing attribute values in several data sets were simply replaced using the same methods by Skalak (1994).

Experimental setup: In this experiment, all data sets were divided into ten approximately equal subsets and was run 10 times each using 10-fold cross validation method similar to (Al-Daoud, 2009). To ensure fairness, the MCS algorithm was directly trained using the accuracy and OAERP metric simultaneously for selecting and discriminating the best solution. The highest value from both metrics through n generated solution will be used for final weight (trained classification model) and tested with test data. We refer these MCS models as MCS_{Acc} and MCS_{OAERP}, respectively. To compute the similarity distance between each training data and the final weight (best solution), the Euclidean distance measurement was employed. The MCS algorithm from (Skalak, 1994). was re-implemented using MATLAB Script version 2009b. All parameters used for this experiment were similar to Skalak (1994) except in the number of generated solution, n. In this experiment, we set n = 500, similar to Bezdek and Kuncheva (2001) which is to ensure that MCS algorithm has enough generated solution to be evaluated during the classification training. This is contrary to implementation of the original MCS algorithm, whereby the total n = 100 used for training process is too small (Skalak, 1994). From this experiment, the expectation was to see that the MCS_{OAERP} model is able to predict better than MCS_{Acc} model when tested with the test data. For evaluation purposes, the average of testing accuracy ($A.TE_{Acc}$) and the average of testing OAERP ($A.TE_{OAERP}$) from ten trial records for each data set were presented and reported for further analysis and comparison.

EXPERIMENTAL RESULTS

Results of experiment 1: Let us consider the first example that focused on balanced class problem.

Example 1: Given a balanced data set containing 50 positive and 50 negative instances (domain Ψ) and two evaluation metrics which are accuracy (Acc) and OAERP are used to discriminate six similar solutions (S) from A to F where Acc = {(A, B, C, D, E, F | A, B, C, D, E, F ϵ Ψ} and OAERP = {(A, B, C, D, E, F) | A, B, C, D, E, F ϵ Ψ}. Assume that all solutions obtained the same total correct predicted instances which are 90 instances as given in Table 3.

Since, this problem is balanced class distribution, intuitively, we can conclude that solution F is better than the other solutions. This is proven by evaluating the value of TP and TN (correctly classified instances) and FP and FN (misclassified instances). In this case, the TP and TN and FP and FN values for F were comparatively balanced as compared to the remaining solutions. Intuitively, all of the above solutions could be ranked as follows: (F>E>D>C>B>A). From this example, the OAERP metric ranks its values similar to our intuition while the accuracy metric was unable to rank its values (undistinguishable) due to poor value produced.

Let us consider another example based on imbalanced class problem.

Example 2: Given an imbalanced data set containing 5 positive and 95 negative instances (domain Ψ) and two evaluation metrics which are accuracy (Acc) and OAERP are used to discriminate six similar solutions (S) from A to F where Acc = {(A, B, C, D, E, F | A, B, C, D, E, F ϵ Ψ} and OAERP = {(A, B, C, D, E, F) | A, B, C, D, E, F ϵ Ψ}. Assume that all solutions obtained the same total correct predicted instances which are 95 instances as given in Table 4.

From all solutions in Table 4, intuitively, the solution A was the poorest solution since it does not has any single positive instance that was correctly classified. In contrast, solution F was the most informative solution since all minority class (positive) instances were correctly classified as compared to other solutions. Recall that when dealing with imbalanced class distribution usually the majority class instances have more influence than the minority class instances. Therefore, if more minority class instances were correctly predicted then the better solution will be obtained. Intuitively, we can rank all solutions according to the degree of informativeness as represented by the TP and TN values as (F>E>D>C>B>A). Similar to Example 1, all solutions are

Table 3: Accuracy against OAERP metric for balanced class problem

S	TP	FP	TN	FN	TC	Acc	OAERP
A	50	10	40	0	90	0.900000	0.890909
B	49	9	41	1	90	0.900000	0.892593
C	48	8	42	2	90	0.900000	0.894340
D	47	7	43	3	90	0.900000	0.896154
E	46	6	44	4	90	0.900000	0.898039
F	45	5	45	5	90	0.900000	0.900000

Table 4: Accuracy against OAERP for imbalanced class problem

S	TP	FP	TN	FN	TC	Acc	OAERP
A	0	0	95	5	95	0.950000	0.950000
B	1	1	94	4	95	0.950000	0.907143
C	2	2	93	3	95	0.950000	0.938889
D	3	3	92	2	95	0.950000	0.940909
E	4	4	91	1	95	0.950000	0.926923
F	5	5	90	0	95	0.950000	0.916667

undistinguishable by the accuracy metric as compared to values produced by OAERP metric. Although, the values produced by OAERP were distinctive and distinguishable, intuitively, it is not aligned with the intuition decision. In this case, the poorest solution A was ranked the highest while the most informative solution F was ranked fifth place. Based on OAERP values, we can rank all solutions as (A>D>C>E>F>B).

Based on two examples discussed, we can assure that the value produced by OAERP metric was distinctive and distinguishable as compared to accuracy value for two different class problems. Unlike the balanced class problem, the value of OAERP metric was not aligned (ill-ranked) with our intuition (well-ranked) for the imbalanced class problem. One might question, does this ill-ranked solution will affect the performance of any stochastic classification algorithm when directly trained by OAERP metric to discriminate and select the best solution for imbalanced class problem? The answer for this question will be testified in the next experiment.

Results of experiment 2: Table 5 shows the average testing results of two MCS models for each data set. Note that the mean of average testing accuracy ($A.TE_{Acc}$) and average testing OAERP ($A.TE_{OAERP}$) values obtained by MCS_{OAERP} model were better than the MCS_{Acc} model. The mean of $A.TE_{Acc}$ and $A.TE_{OAERP}$ value obtained by MCS_{OAERP} model are 0.8464 and 0.8438, respectively while the MCS_{Acc} model obtained 0.8158 and 0.8117, respectively. Overall, the MCS_{OAERP} model shows an outstanding performance against the MCS_{Acc} model whereby the MCS_{OAERP} model has improved the classification accuracy for eight medical data sets. In addition, although the OAERP metric could not show a better decision as compared to human intuition for imbalanced class problem (ill-ranked), this limitation did not impede the MCS_{OAERP} model to achieve better predictive results than the MCS_{Acc} model in classifying nine imbalanced data sets.

To show the significant improvement of the results of MCS_{OAERP} model against MCS_{Acc} model, we perform a paired t-test with 95% confidence level on each medical data set using the ten trial records from each data set. For comparison, we count in how many data sets that one study model is statistically significantly win, tie or loss than another model. As indicated in Table 6, the MCS_{OAERP} model obtained five statistically significant wins against MCS_{Acc} model for both $A.TE_{Acc}$ and $A.TE_{OAERP}$ values based on eight improved results. On top of that, we also perform a paired t-test analysis on overall $A.TE_{Acc}$ and $A.TE_{OAERP}$ values obtained by both MCS models over nine

Table 5: Predictive results from the two MCS models trained by accuracy and OAERP metric for nine medical binary data sets

	Use MCS_{Acc}		Use MCS_{OAERP}	
Data sets	$A.TE_{Acc}$	$A.TE_{OAERP}$	$A.TE_{Acc}$	$A.TE_{OAERP}$
BCO	0.9686	0.9682	0.9629	0.9625
BCD	0.9648	0.9640	0.9737	0.9731
BCP	0.7311	0.7196	0.7776	0.7716
Heart	0.8444	0.8430	0.8778	0.8765
Hepa	0.8063	0.8027	0.8650	0.8640
Liver	0.6235	0.6195	0.6703	0.6690
Pksn	0.8413	0.8324	0.8766	0.8701
Pima	0.7526	0.7501	0.7631	0.7614
SPECTF	0.8095	0.8059	0.8507	0.8463
Mean	0.8158	0.8117	0.8464	0.8438

Table 6: T-test analysis of $\mathrm{MCS}_{\mathrm{OAERP}}$ against $\mathrm{MCS}_{\mathrm{Acc}}$ model using ten trial records from each data set

Data sets	Use $\mathrm{A.TE}_{\mathrm{Acc}}$	Use $\mathrm{A.TE}_{\mathrm{OAERP}}$
BCO	n	n
BCD	n	n
BCP	n	n
Heart	w	w
Hepa	w	w
Liver	w	w
Pksn	w	w
Pima	n	n
SPECTF	w	w
Total (w/n/l)	5/4/0	5/4/0

w: Statistically significant win, n: Statistically not significant, l: Statistically significant loss

medical data sets (Table 5). From this analysis, the $\mathrm{MCS}_{\mathrm{OAERP}}$ model shows significant difference with $\mathrm{MCS}_{\mathrm{Acc}}$ model for both $\mathrm{A.TE}_{\mathrm{Acc}}$ and $\mathrm{A.TE}_{\mathrm{OAERP}}$ values at confidence level of 95% and even 99%, where ρ-value for both analyses are 0.001. Through this experiment, we can conclude that MCS model trained by OAERP metric is better and statistically significant than MCS model trained by accuracy metric.

DISCUSSION

Given the results from the two examples, we proved that by combining the extended precision and extended recall metric into accuracy metric, the results produced have meaningfully changed (distinctive value and distinguishable) as compared to its previous metric values. Although, the values produced by the OAERP metric were distinctive and distinguishable, in Example 2, the values produced by the OAERP metric were not aligned with intuitive decision for imbalanced class problem. We found that the correlation used to determine the Relationship Index (RI) value was unsuitable and not able to portray the performance of each class data. In computing the RI value, we group all precision values together versus a group of recall values which do not portray the relationship among the precision and recall values in each class. Therefore, for the next study, we suggest the correlation given by Lingras and Butz (2007). Lingras and Butz (2007) proved that for two-class problem the precision of one class is correlated to recall of other class and vice versa. Through this correlation, we can construct a different way to compute the RI value which we believed can produce a better value and aligned with human intuitive decision especially for imbalanced class problem.

We also believed that the OAERP metric are well facilitating in MCS searching process, leading towards a better training classification model. The basic idea of MCS algorithm is always heading towards a solution (future state) that is better than the current one (best current state) (Skalak, 1994). In the context of data classification, the future states can be viewed as different classification models. This means the selected best future state can be viewed as selecting the best future model. By employing the OAERP metric as the objective function for MCS algorithm, it shows that the MCS searching process is more likely to choose the better future state due to the advantage of its distinctive and distinguishable value. Unlike the accuracy metric, this metric is easy to get stuck into flat plateau (Table 3, 4) due its less distinctive value (Huang and Ling, 2007). Based on this finding, its prompts us to believe that the OAERP metric can do a better job than the accuracy metric in selecting and discriminating the best solution during the classification training. At the end, this will lead towards a better trained MCS model as compared to the accuracy metric.

On top of that, the experimental results also show that the MCS model trained by OAERP metric performs better than MCS model trained by accuracy metric even if both MCS models are evaluated with testing accuracy value ($A.TE_{Acc}$). Although, contradictory with common intuition in machine learning where a particular model should be optimized by an evaluation metric that it will measure, this finding is consistent with findings from Huang and Ling (2007), Rosset (2004), Skalak *et al.* (2007) and Wu *et al.* (2007).

CONCLUSION

In this study, we have successfully formulated a new evaluation metric called the Optimized Accuracy with Extended Recall-precision (OAERP) based on combination of accuracy, extended recall and extended precision metrics. We have demonstrated that the OAERP metric was better than accuracy metric in terms of distinctness of value produced, discriminating power to choose a better solution and able to build a better trained MCS model. Based on these results, it suggests that OAERP metric should replace the accuracy metric for obtaining a better trained classification model which can lead to better predictive results. For future study, we are planning to extend this proposed metric to solve multi-class problems. Moreover, we are also interested to study and re-design other accuracy-based stochastic classification algorithms to optimize OAERP metric.

ACKNOWLEDGMENT

We wish to thanks the editors and the anonymous reviewers for their comments and suggestions. This research was sponsored by Universiti Malaysia Sarawak (UNIMAS).

REFERENCES

Al-Daoud, E., 2009. A comparison between three neural network models for classification problems. J. Artif. Intell., 2: 56-64.

Al-Shalabi, L., 2011. Knowledge discovery process: Guide lines for new researchers. J. Artifi. Intell., 4: 21-28.

Bezdek, J.C. and L.I. Kuncheva, 2001. Nearest prototype classifier designs: An experimental study. Int. J. Intell. Syst., 6: 1445-1473.

Bradley, A.P., 1997. The use of the area under ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30: 1145-1159.

Buckland, M. and F. Gey, 1994. The relationship between recall and precision. J. Am. Soc. Inform. Sci., 45: 12-19.

Chawla, N.V., N. Japkowicz and A. Kolcz, 2004. Editorial: Special issue on learning from imbalanced data sets. SIGKDD Explorations, 6: 1-6.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recog. Lett., 27: 861-874.

Ferri, C., P.A. Falch and J. Hernandez-Orallo, 2002. Learning decision trees using the area under the ROC curve. Proceedings of the 19th International Conference on Machine Learning, July 2002, Sydney, Australia, pp: 139-146.

Frank, A. and A. Asuncion, 2010. UCI machine learning repository Irvine. University of California, School of Information and Computer Science, USA. http://archive.ics.uci.edu/ml/

Garcia, S. and F. Herrera, 2008. Evolutionary training set selection to optimize C4.5 in imbalance problems. Proceeding of the 8th International Conference on Hybrid Intelligent Systems, September 10-12, 2008, Barcelona, Spain, pp: 567-572.

Hand, D.J. and R.J. Till, 2001. A simple generalization of the area under the ROC curve to multiple class classification problems. Mach. Learn., 45: 171-186.

Huang, J. and C.X. Ling, 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. Knowledge Data Eng., 17: 299-310.

Huang, J. and C.X. Ling, 2007. Constructing new and better evaluation measures for machine learning. Proceedings of the 20th International Joint Conference on Artificial Intelligence, January 6-12, 2007, Hyderabad, India, pp: 859-864.

Lingras, P. and C.J. Butz, 2007. Precision and recall in rough support vector machines. Proceedings of the IEEE International Conference on Granular Computing, November 2-4, 2007, Fremont, CA., pp: 654-654.

MacKay, D.J.C., 2003. Information, Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge, ISBN: 9780521642989, pp: 532-533.

Provost, F. and P. Domingos, 2000. Well-trained PETs: Improving probability estimation trees. CeDER Working Paper #IS-00-04, Stern School of Business, New York University, New York, USA.

Provost, F. and P. Domingos, 2003. Tree induction for probability-based ranking. Mach. Learn., 52: 199-215.

Rakotomamonyj, A., 2004. Optimizing area under ROC with SVMs. Proceedings of the European Conference on Artificial Intelligence Workshop on ROC Curve and Artificial Intelligence, August 22, 2004, Valencia, Spain, pp: 71-80.

Ranawana, R. and V. Palade, 2006. Optimized precision-A new measure for classifier performance evaluation. Proceedings of the IEEE World Congress on Computational Intelligence, July 16-21, 2006, Vancouver, Canada, pp: 2254-2261.

Rosset, S., 2004. Model selection via AUC. Proceedings of the 21st International Conference on Machine Learning, July 4-8, 2004, Banff, Alberta, Canada..

Skalak, D.B., 1994. Prototype and feature selection by sampling and random mutation hill climbing algorithm. Proceedings of the 11th International Conference on Machine Learning, (ICML'94), New Brunswick, New Jersey, pp: 293-301.

Skalak, D.B., A. Niculescu-Mizil and R. Caruana, 2007. Classifier loss under metric uncertainty. Proceedings of the 18th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, September 17-21, 2007, Springer, Heidelberg, pp: 310-322.

Tan, P.N., M. Steibach and V. Kumar, 2006. Introduction to Data Mining. Pearson Addison Wesley, Boston, USA., ISBN: 9780321420527, Pages: 769.

Wilson, S.W., 2001. Mining oblique data with XCS. Adv. Learn. Classifier Syst., 1996: 283-290.

Wu, S., P. Flach and C. Ferri, 2007. An improved model selection heuristic for AUC. Lecture Notes Comput. Sci., 4701: 478-489.

Yan, L., R. Dodier, M.C. Mozer and R. Wolniewicz, 2003. Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. Proceedings of the 20th International Conference on Machine Learning, August 21-24, 2003, Washington, DC., pp: 848-855.