IMPROVING ACCURACY METRIC WITH PRECISION AND RECALL METRICS FOR OPTIMIZING STOCHASTIC CLASSIFIER

Hossin M.[1,2], Sulaiman M.N.[1], Mustpaha N.[1] and Rahmat R.W.[1]

[1]Universiti Putra Malaysia (UPM), Malaysia, mhossin78@gmail.com,
{nasir, norwati, rahmita}@fsktm.upm.edu.my
[2]Universiti Malaysia Sarawak (UNIMAS), Malaysia, hmohamma@fcs.unimas.my

ABSTRACT. All stochastic classifiers attempt to improve their classification performance by constructing an optimized classifier. Typically, all of stochastic classification algorithms employ accuracy metric to discriminate an optimal solution. However, the use of accuracy metric could lead the solution towards the sub-optimal solution due less discriminating power. Moreover, the accuracy metric also unable to perform optimally when dealing with imbalanced class distribution. In this study, we propose a new evaluation metric that combines accuracy metric with the extended precision and recall metrics to negate these detrimental effects. We refer the new evaluation metric as optimized accuracy with recall-precision (OARP). This paper demonstrates that the OARP metric is more discriminating than the accuracy metric and able to perform optimally when dealing with imbalanced class distribution using one simple counter-example. We also demonstrate empirically that a naïve stochastic classification algorithm, which is Monte Carlo Sampling (MCS) algorithm trained with the OARP metric, is able to obtain better predictive results than the one trained with the accuracy and F-Measure metrics. Additionally, the *t*-test analysis also shows a clear advantage of the MCS model trained with the OARP metric over the two selected metrics for almost five medical data sets.

Keywords: optimized classifier, optimal performance, stochastic classification algorithm

INTRODUCTION

Instance selection (IS) is one of the classification methods which aim to reduce the instances as much as possible and simultaneously attempt to achieve the highest possible classification accuracy. From the previous studies, some of the IS methods are developed using stochastic methods such as Monte Carlo (Skalak, 1994), genetic algorithm (Garcia-Pedrajas et al., 2010) and tabu search (Ceveron & Ferri, 2001). In general, these algorithms use the training stage learns from the data and at the same time attempt to optimize the solution by discriminating the optimal solution from the large space of solutions. In order to find the optimal solution, the selection of suitable evaluation metric is essential. According to Ranawana and Palade (2006), to select the suitable evaluation metric for discriminating an optimal solution, the selected evaluation metric must be able to maximize the total number of correct predicted instances in every class. In certain situation, it is hard to build an optimized classifier that can obtain the maximal value for every class. This is because, traditionally, most of the stochastic classification algorithms employ the accuracy rate or the error rate (1-*accuracy*) to discriminate and to select the optimal solution. In (Huang & Ling, 2005; Ranawana & Palade, 2006; Wilson, 1996), they have demonstrated that the simplicity of this accuracy metric could lead to the sub-optimal solutions. For instance, when dealing with imbalanced class instances, it is often happen that the classification model is able to perform extremely well on a large class instances but unfortunately perform poorly on the small class instances. Furthermore, the accuracy metric also exhibits poor discriminating power to discriminate better solution in order to build an optimized classifier (Huang & Ling, 2005, Ling et al., 2003, Rakotomamonyj, 2004).

Based on the drawbacks of the accuracy metric, clearly, this indicates that the main objective of any development of evaluation metric should be able to maximize all class instances in order to build an optimized classifier. Thus, in this study, we are going to propose a new evaluation metric that attempts to improve the accuracy metric. In this study, we are proposing to combine the accuracy metric with the precision and recall metrics. The new evaluation metric is known as an optimized accuracy with recall-precision (OARP) metric.

Precision and recall are two evaluation metrics that are commonly used as the alternative metrics to measure the performance of binary classifiers for two different aspects (Buckland & Gey, 1994). Basically, precision is used to determine the fraction of positive instances that are correctly predicted in a positive class, while recall measures the fraction of positive instances being correctly classified over the total of positive instances. However, it is not easy to apply both precision and recall metrics separately because it will turn the selection and discrimination processes more difficult due to multiple comparisons. In fact, this strategy can lead to the sub-optimal solution especially when the classifier attempts to maximize both metrics simultaneously. Moreover, the conventional precision and recall metrics are not suitable to be employed for the combination process with the accuracy metric. This is because both metrics only measure one class of instances (positive class). This is somewhat against the ideal idea of formulating the best evaluation metric as aforesaid, which is must be able to maximal the correct predicted instances for every class. To resolve this limitation, the extended precision and recall metrics proposed by (Lingras & Butz, 2007) were suggested for the combination. The main justification is that every class instance should be able to be measured individually using both metrics.

In this paper, we will show that our newly constructed evaluation metric will improve the conventional accuracy metric using one counter-example in terms of discriminatory and perform optimally when dealing with imbalanced class distribution. To prove this theoretical evidence, we demonstrate empirically that the OARP metric is better than conventional accuracy metric using a naïve stochastic classification in classifying five medical data sets that obtained from UCI Machine Learning Repository (Frank & Asuncion, 2009). From this experiment, the expectation is to see that the naïve stochastic algorithm trained by the OARP metric will produce better predictive result than the one trained by the accuracy metric.

OPTIMIZED ACCURACY WITH PRECISION AND RECALL (OARP)

As aforesaid, the purpose of this study is to improve the accuracy metric by combining the accuracy metric with the extended precision and recall metrics. In order to combine these metrics into a singular form of metric, we have adopted two important formulas from (Ranawana & Palade, 2006), which are the Relationship Index (RI) and OP. Due to limited pages, the details of these reference metrics can be found in (Lingras & Butz, 2007; Ranawana & Palade, 2006). The combination process involves two-step efforts, whereby first we have to find a suitable way to employ the RI formula and next is to identify the best approach to adopt the OP formula in order to improve the accuracy metric.

As proved by (Lingras & Butz, 2007), for two-class problem, the extended precision value in a particular class is proportional to the extended recall values of the other class and vice versa. From this correlation, the RI formula can be implemented. To employ the RI formula, the precision and recall from different classes were paired together (p_1, r_2), (p_2, r_1) based on the correlation given in (Lingras & Butz, 2007). At this point, the aim is to minimize the value of

$|p_1\text{-}r_2|$ and $|p_2\text{-}r_1|$, and maximize the value of p_1+r_2 and p_2+r_1. Hence, we define the RI for both correlations as stated in Eq. (1) and (2).

$$RI_1 = \frac{p_1 - r_2}{p_1 + r_2} \tag{1}$$

$$RI_2 = \frac{p_2 - r_1}{p_2 + r_1} \tag{2}$$

However, these individual RI values are still pointless and could not be applied directly to calculate the value of new evaluation metric. Thus, to resolve this problem, we compute the average of total RI (AVRI) as shown in Eq. (3) to formulate the new evaluation metric.

$$AVRI = \frac{RI_1 + RI_2}{2} \tag{3}$$

As mentioned earlier, the use of accuracy value alone could lead the searching process to the sub-optimal solutions mainly due to its less discriminative power and inability to deal with imbalanced class distribution. Such drawbacks motivate us to combine the beneficial properties of AVRI with the accuracy metric. With this combination, we expect the new evaluation metric is able to produce better value (more discriminating) than the accuracy metric and at the same time remain relatively stable when dealing with imbalanced class distribution. The new evaluation metric is called the optimized accuracy with recall-precision (OARP) metric. The computation of this OARP metric is defined in Eq. (14).

$$OARP = Acc - AVRI \tag{4}$$

However, during the computation of this new evaluation metric, we noticed that the value of OARP may deviate too far from the accuracy value especially when the value of AVRI is larger than accuracy value. Therefore, we proposed to resize the AVRI value into a small value before computing the OARP metric. To resize the AVRI value, we employed the decimal scaling method to normalize the AVRI value as shown in Eq. (5).

$$AVRI_{new\_val} = \frac{AVRI_{old\_val}}{10^x} \tag{5}$$

where x is the smallest integer such that max ($|AVRI_{new\_val}|$) < 1. In this study, we set the x=1 for the entire experiments. By resizing the AVRI value, we found that the OARP value is comparatively close to the accuracy value as shown in the next sub-section. At the end, the objective of OARP metric is to optimize the classifier performance. A high OARP value entails a low value of AVRI which indicates a better generated solution has been produced. We also noticed that via this new evaluation metric, the OARP value is always less than the accuracy value ($OARP < Acc$). The OARP value will only equal to the accuracy value ($OARP=Acc$) when the AVRI value is equivalent to 0 ($AVRI=0$), which indicates a perfect training classification result (100%).

EMPIRICAL VERIFICATION

In this particular section, two types of empirical verification have been conducted in order to verify the advantage of OARP metric. Firstly, we compare the OARP metric with the conventional accuracy metric using one simple counter-example. Secondly, we empirically compare the OARP metric with the accuracy and F-Measure metrics for selecting and discriminating five medical data sets using a naïve stochastic classification algorithm.

OARP vs. Accuracy using Counter-examples

In this particular sub-section, we attempt to demonstrate that the OARP metric is better than the accuracy metric using the following counter-example. Let us consider counter-example as shown in Table 1 that focused on imbalanced class distribution. In this counter-example, the accuracy metric could not distinguished whether *a* or *b* is better, while the OARP metric otherwise. Intuitively, we can conclude that *b* is better than *a*. This is because, *b* is able to predict correctly all the minority class instances if compared to *a*. Clearly, *a* is poor since no single instance from minority class instances is correctly predicted by *a* (non-informative output for the minority class). Hence, we can conclude that the result obtained by the OARP metric is similar to intuitive decision and clearly better than the accuracy metric in discriminating the optimal solution. On top of that, the counter-example in Table 1 also shows that the accuracy metric could not work optimally when dealing with imbalanced class distribution.

Table 1. Accuracy vs. OARP for imbalanced data set (95:5)

s	tp	fp	tn	fn	TC	Accuracy	OARP
a	95	5	0	0	95	0.950000	0.850000
b	90	0	5	5	95	0.950000	0.934545

Note: tp-true positive, fp-false positive, tn-true negative, fn-false negative, TCC-total correct classified

Real Data Sets

As we established in the previous section, it is not enough to claim that the OARP metric is better than accuracy metric using one simple counter-example. Through the counter-example, we only can demonstrate a very little evidence in order to prove that the OARP metric is really better than the accuracy metric. Thus, in this particular section, we are going to demonstrate the generalization capability of the OARP metric using real world application data sets. Instead of accuracy metric, we add another existing metric that is F-measure (van Rijsbergen, 1979) to compare with the OARP metric. F-measure is chosen to represents the conventional precision and recall metrics. As aforesaid, it is hard to apply the precision and recall metrics separately, thus, F-measure is the best way to represents these two metrics. In fact, F-measure is proven to be the more favorable evaluation metric for evaluating the imbalanced class distribution (Joshi, 2002).

Experimental Setup. For the purpose of comparison and evaluation on the capability of OARP metric against the accuracy and F-measure metrics, five medical data sets from UCI Machine Learning Repository (Frank & Asuncion, 2010) were selected. The brief descriptions about these selected data sets are summarized in Table 2.

Table 2: Brief description of each medical data set.

Dataset	No. of Instances	No. of Attributes	Missing Value	Class Distribution
Breast-cancer	699	9	Yes	IM
Heart270	270	13	No	IM
Hepatitis	155	19	Yes	IM
Liver	345	6	No	IM
Pima-diabetes	768	8	No	IM

All data sets have been normalized within the range of [0, 1] using min-max normalization. Normalized data is essential to speed up the matching process for each attribute and prevent any attribute variables from dominating the analysis (Al-Shalabi et al., 2006). All missing attribute values in several data sets were simply replaced with median value for numeric value and mode value for symbolic value of that particular attribute across all instances. In this study, all data sets were divided into ten approximately equal subsets using 10-fold cross validation method similar to (Garcia-Pedrajas et al., 2010). Each data set was run for 10 times.

In this experiment, all of selected data sets were trained using a naïve stochastic classification algorithm which is Monte Carlo Sampling algorithm (Skalak, 1994). This algorithm combines simple stochastic method (random search) and instance selection strategy. There are two main reasons this algorithm is selected. Firstly, this algorithm simply applies accuracy metric to discriminate the optimal solution during the training phase. Secondly, this algorithm is aligned with the purpose of this study which is to optimize the stochastic classification algorithm. To compute the similarity distance between each training instance and prototype solution (each class has one representative instance), the Euclidean distance measurement is employed. The MCS algorithm was re-implemented using MATLAB Script version 2009b. To ensure fair experiment, the MCS algorithm was trained simultaneously using the accuracy, F-Measure and OARP metrics for selecting and discriminating the optimal solution. For simplicity, we refer these four MCS models as $\mathrm{MCS}_{\mathrm{Acc}}$, $\mathrm{MCS}_{\mathrm{FM}}$ and $\mathrm{MCS}_{\mathrm{OARP}}$ respectively. All parameters used for this experiment are similar to (Skalak, 1994) except in the number of generated solution, *n*. In this experiment, we employed *n*=500 similar to (Bezdek & Kuncheva, 2002). From this experiment, the expectation is to see that the $\mathrm{MCS}_{\mathrm{OARP}}$ is able to predict better than the model optimized by the $\mathrm{MCS}_{\mathrm{Acc}}$ and $\mathrm{MCS}_{\mathrm{FM}}$. For evaluation purposes, the average of testing accuracy ($\mathrm{Test}_{\mathrm{Acc}}$) will be used for further analysis and comparison.

Experimental Results. Table 3 shows the average testing accuracy for each data set based on each MCS model. From Table 3, we can see that the average testing accuracy obtained by $\mathrm{MCS}_{\mathrm{OARP}}$ is better than the $\mathrm{MCS}_{\mathrm{Acc}}$ and $\mathrm{MCS}_{\mathrm{FM}}$ models. The average testing accuracy obtained by $\mathrm{MCS}_{\mathrm{OARP}}$ model is 0.8542 while the $\mathrm{MCS}_{\mathrm{Acc}}$ and $\mathrm{MCS}_{\mathrm{FM}}$ models obtained 0.8186 and 0.7806 respectively for all five medical data sets. On top of that, the $\mathrm{MCS}_{\mathrm{OARP}}$ model has improved the classification performance in all data sets if compared to $\mathrm{MCS}_{\mathrm{Acc}}$ and $\mathrm{MCS}_{\mathrm{FM}}$ models.

To verify this outstanding performance, we perform a paired *t*-test with 95% confidence level on each medical data set by using the ten trial records from each data set. The summary result of this comparison is listed in Table 4. As indicated in Table 4, the $\mathrm{MCS}_{\mathrm{OARP}}$ model obtained four statistically significant wins against both $\mathrm{MCS}_{\mathrm{Acc}}$ and $\mathrm{MCS}_{\mathrm{FM}}$ models. Meanwhile only one data set (Heart270) shows no significant differences from both comparisons.

Table 3: Average testing accuracy for both MCS models.

Data set	Use $\mathrm{MCS}_{\mathrm{Acc}}$	Use $\mathrm{MCS}_{\mathrm{FM}}$	Use $\mathrm{MCS}_{\mathrm{OARP}}$
	$\mathrm{Test}_{\mathrm{Acc}}$	$\mathrm{Test}_{\mathrm{Acc}}$	$\mathrm{Test}_{\mathrm{Acc}}$
Breast-Cancer	0.9700	0.9685	*0.9814*
Heart270	0.8704	0.8556	*0.8778*
Hepatitis	0.8454	0.8183	*0.8900*
Liver	0.6468	0.5302	*0.7160*
Pima-diabetes	0.7513	0.7305	*0.8060*
Average	0.8168	0.7806	**0.8542**

Table 4. Comparison summary of the *t*-test analysis based on ten trial records for each medical data set.

Data set	$\mathrm{MCS}_{\mathrm{OARP}}$ vs. $\mathrm{MCS}_{\mathrm{Acc}}$	$\mathrm{MCS}_{\mathrm{OARP}}$ vs. $\mathrm{MCS}_{\mathrm{FM}}$
Breast-Cancer	*Ssw*	*ssw*
Heart270	*Sns*	*sns*
Hepatitis	*Ssw*	*ssw*
Liver	*ssw*	*ssw*
Pima-diabetes	*ssw*	*ssw*

Note: *ssw*-statistically significant win, *ssl*-statistically significant loss, *sns*-statistically not significant

CONCLUSION AND FUTURE WORKS

In this paper, we have proposed a new evaluation metric called the Optimized Accuracy with Recall-Precision (OARP) based on combination of three existing metrics, which are the accuracy, and the extended recall and precision metrics. Theoretically, we have proved that our newly constructed evaluation metric is better than conventional accuracy metric using a simple counter-example. From this counter-example, we have showed that the OARP metric is more discriminating than accuracy metric. More importantly, the OARP also shows that it can work optimally when dealing with the imbalanced class distribution. To support our theoretical evidence, we have compared experimentally the OARP metric against the accuracy metric using five medical data sets. In this experiment, we have added the F-Measure metric for representing the conventional precision and recall metrics. Interestingly, the naïve stochastic classification algorithm, which is Monte Carlo Sampling (MCS) algorithm optimized by the OARP metric has outperformed and statistically significant than the MCS algorithm optimized by the accuracy and F-Measure metrics. This indicates that the OARP metric is more likely to choose an optimal solution in order to build an optimized stochastic classifier. For the future work, we are planning to extend this new evaluation metric, OARP for solving multi-class problems. Moreover, we are also interested to verify the advantage of the OARP metric using a statistical consistency and discriminatory analysis proposed by Huang and Ling (2005).

REFERENCES

Al-Shalabi, L., Shaaban, Z., & Kasasbeh, B. (2006). Data Mining: A Preprocessing Engine, *Journal of Computer Science,* 2(9),735-739.

Bezdek, C. J., & Kuncheva, L.I. (2001). Nearest Prototype Classifier Designs: An Experimental Study. *International Journal of Intelligent Systems*, 6, 1445-1473.

Buckland, M., & Gey, F. (1994). The Relationship between Recall and Precision. *Journal of the American Society for Information Science*, 45(1), 12-19.

Cerveron V., & Ferri, F.J. (2001). Another move toward the minimum consistent subset: A Tabu Search approach to the Condensed Nearest Neighbor rule, *IEEE Transactions on Systems, man, and Cybernetics-Part B: Cybernetics* 31(3), 408-413.

Frank, A., & Asuncion, A. (2009, October 5). UCI Machine Learning Repository: Center for Machine Learning and Intelligent Systems. Retrieved from http://archive.ics.uci.edu/ml

Garcia-Pedrajas, N., Romero del Castillo, J.A., & Ortiz-Boyer, D. (2010). A cooperative coevolutionary algorithm for instance selection for instance-based learning. *Machine Learning*, 78, 381-420.

Huang, J., & Ling, C.X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3), 299-310.

Joshi, M.V. (2002). On evaluating performance of classifiers for rare classes. In *Proceedings of ICDM'02* (pp. 641-644). Maebashi, Japan.

Ling, C.X., Huang, J., & Zhang, H. (2003). AUC: A Statistically Consistent and More Discriminating Measure than Accuracy. In *Proceedings of 18th International Conference on Artificial Intelligence (IJCAI-2003)* (pp. 519-526).

Lingras, P., & Butz, C.J. (2007). Precision and Recall in Rough Support Vector Machines. In *2007 IEEE International Conference on Granular Computing (GRC 2007)* (pp. 654). San Jose, California.

Rakotomamonyj, A. (2004). Optimizing area under ROC with SVMs. In J. Hernandez-Orallo, C. Ferri, N. Lachiche, and P. Flach (Eds.), *Proceedings of the European Conference on Artificial Intelligence Workshop on ROC Curve and Artificial Intelligence (ROCAI 2004)* (pp. 71-80). Valencia, Spain.

Ranawana, R., & Palade, V. (2006). Optimized Precision - A New Measure for Classifier Performance Evaluation. In *Proceedings of the IEEE World Congress on Computational Intelligence* (pp. 2254-2261). Vancouver, Canada.

Skalak, D.B. (1994). Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithm, In W.W. Cohen, & H. Hirsh (Eds.), *International Conference on Machine Learning* (pp. 293-301). New Brunswick, NJ: Morgan-Kaufmann.

van Rijsbergen, C.J. (1979). *Information Retrieval*. Butterworth's London.

Wilson, S.W. (2001). Mining oblique data with XCS. *Lecture Notes in Computer Science*, 1996, 158-176. Springer-Verlag.