LEC.1 Paper Work

Submitted by: Kavya Sharma(15MI405)

A Review on Evaluation Metrics for Data Classification Evaluations

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 and 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.

Data classification can be divided into binary, multiclass and multi-labelled classification. Binary and multiclass classification focuses on the evaluation metrics for evaluating the effectiveness of classifiers. The evaluation metric can be categorized into three types, which are threshold, probability and ranking metric. These metrics can be employed into three different evaluation applications:

1. To measure and summarize the quality of trained classifier when tested with the unseen data. Accuracy or error rate is used to evaluate the generalization ability of classifiers. Accuracy refers to the total of instances that are correctly predicted by the trained classifier when tested with the unseen data.
2. An evaluator for model selection. In this case, the task is to determine the best Classifier among different types of trained classifiers which focus on the best future performance when tested with unseen data.
3. As a discriminator to discriminate and select the optimal solution among all generated solutions during the classification training.

This paper emphasizes on the third application of evaluation metrics for Prototype Selection (PS) Classifiers. It is 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. This algorithms begin with constructing and searching a fixed number of prototypes.

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. The accuracy has a number of limitations. It could lead to suboptimal solutions and also exhibits poor discriminating values to discriminate better solution in order to build an optimized classifier.

Discriminator Metrics

Threshold Types of Discriminator Metrics:

For binary classification problems, the discrimination evaluation of the best solution during the classification training can be defined based on confusion matrix as shown in Table 1.

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)

Accuracy 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. Advantages of accuracy or error rate:

1. This is easy to compute with less complexity.
2. Applicable for multi-class and multi-label problems.
3. Easy-to-use scoring.
4. Easy to understand by human.

Limitations of Accuracy metrics:

1. It produces less distinctive and less discriminable values.
2. It is powerless in terms of informativeness and less favour towards minority class instances.

Table 2. Threshold Metrics for Classification Evaluations

Metrics	**Formula**	**Evaluation Focus**
Accuracy (acc)	(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)	(fp + fn)/(tp + fp + tn + fn)	Misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated.
Sensitivity (sn)	tp/(tn+fn)	This metric is used to measure the fraction of positive patterns that are correctly classified
Specificity (sp)	tn/(tn+fn)	This metric is used to measure the fraction of negative patterns that are correctly classified.
Precision (p)	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)	tp/(tn+fn)	Recall is used to measure the fraction of positive patterns that are correctly classified
F-Measure (FM)	(2*p*r)/(p+r)	This metric represents the harmonic mean between recall and precision values.
Averaged Accuracy	$\sum_{i=1}^{s} \frac{tp_i + tn_i}{tp_i + fn_i \; fp_i}$	The average effectiveness of all classes.
Averaged Error Rate	$\sum_{i=1}^{s} \frac{fp_i + fn_i}{tp_i + fn_i + fp_i}$	The average error rate of all classes.

Averaged F-Measure	$\frac{2 * p_M * r_M}{p_M + r_M}$	The average of per-class F-measure.
Averaged Recall	$\sum_{i=1}^{s} tpi/(tp_i + fn_i)$	The average of per-class recall.

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.

Mean Square Error (MSE)

Supervised LVQ(learning vector quantization) uses MSE to evaluate its performances during the classification training. MSE measures the difference between the predicted solutions and desired solutions. The smaller MSE value is required in order to obtain a better trained supervised LVQ. The MSE is defined as below:

$$\text{MSE}= (1/n) \sum_{j=1} (Pj - Aj)^2$$

where P_j is the predicted value of instance j, A_j is real target value of instance and n is the total number of i.

Limitations of MSE:

1. It does not provide the trade-off information between class data. This may lead the discrimination process to select the sub- optimal solution.
2. It is really dependent on the weight initialization process. In extremely imbalanced class problem, if the initial weights are not proper selected this may lead the discrimination process ends up with sub-optimal solution.

Area under the ROC Curve (AUC)

AUC is one of the popular ranking type metrics. The AUC was used to construct an optimized learning model and also for comparing learning algorithms. AUC value reflects the overall ranking performance of a classifier. For two-class problem, the AUC value can be calculated as below:

$$\text{AUC} = \frac{Sp - np(nn + 1)/2}{np\ nn}$$

where, S_p is the sum of the all positive examples ranked, n_p and n_n denote the number of positive and negative examples respectively.

Advantages of AUC:

The AUC was proven theoretically and empirically better than the accuracy metric for evaluating the classifier performance and discriminating an optimal solution during the classification training.

Limitations of AUC:
The computational cost of AUC is high especially for discriminating a volume of generated solutions of multiclass problems.

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. 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

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

where *acc* is the accuracy score

sp and *sn* denotes specificity and sensitivity score respectively.

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

1. The Optimized Accuracy with Extended Recall-Precision version 1 (OAERP1)
2. Optimized Accuracy with Extended Recall-Precision version 2 (OAERP2).

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).

$$\text{OAERP} = \text{acc} - \text{RI}_1$$

$$RI_1 = \frac{|ep1 + ep2| - |er1 + er2|}{|ep1 + ep2| + |er1 + er2|}$$

$$OAERP = acc - RI_2$$

$$RI_2 = \frac{|ep_1 - er_2|}{|ep_1 + er_2|} - \frac{|ep_2 - er_1|}{|ep_2 + ec_1|}$$

ep and er represent extended precision and extended recal respectively and the numbering denotes class 1 (positive class) and class 2 (negative class).

Limitation of Hybrid Metrics:

It is only limited for discriminating and evaluating the binary classification problems.

Important Factors in Constructing Metrics

The lists of important factors in designing and constructing a new metric or choosing the suitable metric for discriminating the optimal solution of PS classification algorithms 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. 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.We need 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. 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.
4. Informativeness: Another drawback of many current metrics is there is no trade-off information between classes. For example, the most popular metric accuracy could not discriminate the good and bad 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.

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.

5. Favour towards the minority class: 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. For example in Table 5, any good metric or discriminator should rank the solution as follows: $a_6 \rightarrow a_5 \rightarrow a_4 \rightarrow a_3 \rightarrow a_2 \rightarrow a_1$.

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.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.

sol	tp	fp	tn	fn	total	accuracy
a1	0	5	95	0	95	0.9500
a2	5	0	90	5	95	0.9500

Table3. Informativeness Analysis for Binary Classification Problem using Imbalanced Class Distribution (5:95) with Two Equivalent Solutions

sol	tp	fp	tn	fn	total	accuracy
a1	0	0	95	5	95	0.9500
a2	5	6	89	0	94	0.9400

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

sol	tp	fp	tn	fn	total
a1	0	0	9995	5	9995
a2	1	1	9994	4	9995
a3	2	2	9993	3	9995
a4	3	3	9992	2	9995
a5	4	4	9991	1	9995
a6	5	5	9990	0	9995

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

CONCLUSION

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.

REFERENCE: Hossin, M. and Sulaiman, M.N. “A review on evaluation metrics for data classification evaluations”International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.2, March 2015