Gaussian Process Learning from Order Relationships Using Expectation Propagation ................................................................. 605
Ruixuan Wang and Stephen James McKenna

Feature Ranking Based on Decision Border ......................................................... 609
Claudia Diamantini, Alberto Gemelli, and Domenico Potena

Three-layer Spatial Sparse Coding for Image Classification ................................. 613
Dengxin Dai, Wen Yang, and Tianfu Wu

Theoretical Analysis of a Performance Measure for Imbalanced Data ................. 617
V. García, Ramon A. Mollineda, and J. Salvador Sánchez

Cluster Preserving Embedding ........................................................................... 621
Yubin Zhan and Jianping Yin

Color Image Analysis by Quaternion Zernike Moments ....................................... 625
Beijing Chen, Huazhong Shu, Hui Zhang, Gang Chen, and Limin Luo

Topic-Sensitive Tag Ranking ............................................................................... 629
Yan'an Jin, Ruixuan Li, Zhengding Lu, Kunmei Wen, and Xiwu Gu

Water Reflection Detection Using a Flip Invariant Shape Detector ..................... 633
Hua Zhang, Xiaojie Guo, and Xiaochun Cao

CDP Mixture Models for Data Clustering .......................................................... 637
Yangfeng Ji, Tong Lin, and Hongbin Zha

A Simple Approach to Find the Best Wavelet Basis in Classification Problems .......... 641
Farhad Faradji, Rabab K. Ward, and Gary E. Birch

Learning Probabilistic Models of Contours ......................................................... 645
Laure Amate and Maria João Rendas

Local Sparse Representation Based Classification .............................................. 649
Chun-Guang Li, Jun Guo, and Hong-Gang Zhang

Manifold Modeling with Learned Distance in Random Projection Space
for Face Recognition ........................................................................................... 653
Grigorios Tsagkatakis and Andreas Savakis

Part Detection, Description and Selection Based on Hidden Conditional
Random Fields ....................................................................................................... 657
Wenhao Lu, Shengjin Wang, and Xiaoqing Ding

Boosting Bayesian MAP Classification ................................................................. 661
Paolo Piro, Richard Nock, Frank Nielsen, and Michel Barlaud

Weighting of the k-Nearest-Neighbors ................................................................. 666
Konstantin Chernoff and Mads Nielsen

Learning Sparse Face Features: Application to Face Verification ....................... 670
Pierre Buyssens and Marinette Revenu

Image Feature Extraction Using 2D Mel-Cepstrum ............................................ 674
Serdar Çakir and A. Enis Çetin

Entropy Estimation and Multi-Dimensional Scale Saliency ................................ 678
Pablo Suau and Francisco Escolano
Theoretical Analysis of a Performance Measure for Imbalanced Data

V. García, R.A. Mollineda and J.S. Sánchez
Institute of New Imaging Technologies - Dept. Llenguatges i Sistemes Informàtics, Universitat Jaume I
Av. Sos Baynat s/n, 12071 Castelló de la Plana, Spain
{Jimenezv,mollined,sanchez}@uji.es

Abstract

This paper analyzes a generalization of a new metric to evaluate the classification performance in imbalanced domains, combining some estimate of the overall accuracy with a plain index about how dominant the class with the highest individual accuracy is. A theoretical analysis shows the merits of this metric when compared to other well-known measures.

1. Introduction

Most of learning methods assume that the classes of the problem share similar prior probabilities. However, in many real-world tasks the ratios of prior probabilities between classes are significantly skewed. This is known as the imbalance problem [10]. A two-class data set is said to be imbalanced when one of the classes is heavily under-represented as regards the other class [6]. Because of examples of the minority and majority classes usually represent the presence and absence of rare cases, respectively, they are also known as positive and negative examples.

As claimed by many authors [2, 3, 8], the use of plain accuracy (error) rates to evaluate the classification performance in imbalanced domains might produce misleading conclusions, since they do not take misclassification costs into account, are strongly biased to favor the majority class, and are sensitive to class skews.

Most of alternative metrics are formulated as combinations of accuracy (error) rates measured separately on each class, to alleviate biased results. Nevertheless, none of these show up how dominant the accuracy of an individual class is over another, nor distinguish the contribution of each class to the overall performance.

2. Performance evaluation measures

Traditionally, classification accuracy (Acc) and/or error rates have been the standard metrics used to estimate the performance of learning systems. For a two-class problem, they can be easily derived from a $2 \times 2$ confusion matrix as that given in Table 1.

<table>
<thead>
<tr>
<th>Predicted positive</th>
<th>Predicted negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive class</td>
<td>True Positive (TP)</td>
</tr>
<tr>
<td>Negative class</td>
<td>False Positive (FP)</td>
</tr>
</tbody>
</table>

However, empirical and theoretical evidences show that these measures are biased with respect to data imbalance and proportions of correct and incorrect classifications. These shortcomings have motivated a search for new metrics based on simple indexes, such as the true positive rate (TPrate) and the true negative rate (TNrate). The TPrate (TNrate) is the percentage of positive (negative) examples correctly classified.

One of the most widely-used evaluation methods in the context of class imbalance is the ROC curve, which is a tool for visualizing and selecting classifiers based on their trade-offs between benefits (true positives) and costs (false positives). A quantitative representation of a ROC curve is the area under it (AUC) [1]. For just one run of a classifier, the AUC can be computed as [9]

$$AUC = (TPrate + TNrate)/2.$$
Kubat et al. [7] use the geometric mean of accuracies measured separately on each class, with the aim of maximizing the accuracies of both classes while keeping them balanced, \(G\text{mean} = \sqrt{T\text{Pre} \cdot T\text{Nrate}}\).

Both AUC and Gmean minimize the negative influence of skewed distributions of classes, but they do not show up the contribution of each class to the overall performance, nor which is the prevalent class. This means that different combinations of TPre and TNrate may produce the same result for those metrics.

Recently, Ranawana and Palade [8] introduced the optimized precision, which can be computed as,

\[
OP = \frac{Acc - |TNrate - TPre|}{TNrate + TPre} \tag{1}
\]

This represents the difference between the global accuracy and a second term that computes how balanced both class accuracies are. High OP values require high global accuracy and well-balanced class accuracies. However, OP can be strongly affected by the biased influence of the global accuracy.

### 3. Generalizing a new performance metric

This section provides a generalization of a primary index reported in [4], named Index of Balanced Accuracy (IBA). The main purpose of the generalized IBA will be to weight a measure suitable to evaluate the performance in imbalanced domains. The weighting factor will aim at favoring those results with better classification rates on the minority class.

The generalized IBA can be formulated as follows:

\[
IBA_\alpha(M) = (1 + \alpha \cdot Dom) \cdot M \tag{2}
\]

where \((1 + \alpha \cdot Dom)\) is the weighting factor and \(M\) represents any performance metric.

The \(Dom\) term, called dominance, is defined as \(Dom = TPre - TNrate\) within the range \([-1, +1]\), and it is here used to estimate the relationship between the TPre and TNrate. The closer the dominance is to 0, the more balanced both individual rates are. If \(TPre > TNrate\), then \(Dom > 0\); otherwise, \(Dom < 0\).

The value of \(Dom\) is weighted by \(\alpha \geq 0\) to reduce its influence on the result of the particular metric \(M\). Thus the weighting factor in Eq. 2 is within the range \([1 - \alpha, 1 + \alpha]\). Note that if \(\alpha = 0\) or \(TPre = TNrate\), the IBA\(_\alpha\) turns into the measure \(M\). In practice, one should select a value of \(\alpha\) depending on the metric used.

#### 3.1. Formulating IBA\(_\alpha\) with Gmean

As a representative example, this paper will use Gmean because this is a suitable, well-known performance measure for class imbalanced problems. Hence IBA\(_\alpha\) can now be rewritten in terms of Gmean as:

\[
IBA_\alpha(G\text{mean}) = (1 + \alpha \cdot Dom) \cdot G\text{mean} \tag{3}
\]

Since \(\alpha\) will depend on the metric \(M\), the following study is devoted to empirically set an appropriate value of \(\alpha\) for the particular case of IBA\(_\alpha\)(Gmean). Also, this example will allow to clear up the behavior differences of IBA\(_\alpha\) with respect to other metrics.

Let \(f(\theta)\) be a classifier that depends on a set of parameters \(\theta\). Suppose that \(\theta\) should be optimized so that \(f(\theta)\) can discriminate between the two classes of a particular imbalanced problem (with a ratio 1:10). Let \(T\) and \(V\) be the training and validation sets, respectively. During learning, seven possible configurations \((\theta_1, \theta_2, \ldots, \theta_7)\) have been obtained from \(T\), and then the corresponding classifiers \(f(\theta_i)\) have been run over \(V\). Table 2 reports the results of some measures used to evaluate each classifier \(f(\theta_i)\). The last step in learning should be to pick up the best configuration \(\theta^*\) according to the performance measure adopted.

<table>
<thead>
<tr>
<th>(TPre)</th>
<th>(TNrate)</th>
<th>Acc</th>
<th>Gmean</th>
<th>AUC</th>
<th>OP</th>
<th>IBA(_{0.05})</th>
<th>IBA(_{0.1})</th>
<th>IBA(_{0.2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\theta_1)</td>
<td>0.350</td>
<td>0.950</td>
<td>0.914</td>
<td>0.723</td>
<td>0.750</td>
<td>0.647</td>
<td>0.708</td>
<td>0.694</td>
</tr>
<tr>
<td>(\theta_2)</td>
<td>0.650</td>
<td>0.850</td>
<td>0.832</td>
<td>0.743</td>
<td>0.750</td>
<td>0.698</td>
<td>0.736</td>
<td>0.728</td>
</tr>
<tr>
<td>(\theta_3)</td>
<td>0.700</td>
<td>0.800</td>
<td>0.791</td>
<td>0.748</td>
<td>0.750</td>
<td>0.724</td>
<td>0.745</td>
<td>0.741</td>
</tr>
<tr>
<td>(\theta_4)</td>
<td>0.750</td>
<td>0.750</td>
<td>0.750</td>
<td>0.750</td>
<td>0.750</td>
<td>0.750</td>
<td>0.750</td>
<td>0.750</td>
</tr>
<tr>
<td>(\theta_5)</td>
<td>0.800</td>
<td>0.700</td>
<td>0.709</td>
<td>0.748</td>
<td>0.750</td>
<td>0.642</td>
<td>0.752</td>
<td>0.756</td>
</tr>
<tr>
<td>(\theta_6)</td>
<td>0.850</td>
<td>0.650</td>
<td>0.668</td>
<td>0.743</td>
<td>0.750</td>
<td>0.535</td>
<td>0.751</td>
<td>0.758</td>
</tr>
<tr>
<td>(\theta_7)</td>
<td>0.950</td>
<td>0.550</td>
<td>0.586</td>
<td>0.723</td>
<td>0.750</td>
<td>0.320</td>
<td>0.737</td>
<td>0.752</td>
</tr>
</tbody>
</table>

Note that configurations \(\theta_1\) and \(\theta_7\) correspond to cases with a clearly biased behavior, whereas \(\theta_4\) produces a perfect balance between TPre and TNrate. The rest of configurations \(\theta_2, \theta_3, \theta_5\) and \(\theta_6\) produce less differences between TPre and TNrate.

For this example, AUC is of no value at all since all configurations give the same value. Accuracy would select the biased \(\theta_1\) because it strongly depends on the majority class rate. Both Gmean and OP suggest the most balanced configurations \((\theta_3, \theta_4, \theta_5)\), ignoring the fact that the minority class is usually the most important. While Gmean does not distinguish between \(\theta_3\) and \(\theta_5\), OP would prefer \(\theta_3\) rather than \(\theta_5\) because its computation is affected by the accuracy. These drawbacks can be overcome when using the IBA\(_\alpha\) measure by appropriately tuning the parameter \(\alpha\). One can see that IBA\(_{0.05}\) and IBA\(_{0.1}\) select \(\theta_5\) or \(\theta_6\), which correspond to the moderate cases with the highest TPre.

Results of IBA\(_{0.2}\) show a biased tendency of IBA\(_\alpha\) towards TPre for high and moderate values of \(\alpha\). This effect is due to the strong influence of \(Dom\) on IBA\(_\alpha\), what justifies the need of \(\alpha\) to weight its importance. This study suggests that the use of small values of \(\alpha\) allow to correct this effect and thus, we propose \(\alpha = \ldots\)
0.05 for the calculation of IBA_\alpha(Gmean).

4. The theoretical analysis of IBA_\alpha

Two theoretical studies are performed to explore the possible advantages of IBA_\alpha (with \( \alpha = 0.05 \) and \( M = Gmean \)) over other metrics. One computes Pearson correlation coefficients in order to devise how IBA_\alpha is correlated with other metrics that might be deemed as good or bad choices to tackle the imbalance. The second study analyzes how sensitive the metrics are under different types of changes to the confusion matrix.

4.1. Correlation analysis

Five collections of classifier output tuples based on different imbalance ratios were generated as in [5]. A classifier output tuple consists of a list of \( n \) numeric values between 0 and 1 which represent, for \( n \) hypothetical samples, the probabilities of belonging to the positive class (classifier outputs). All tuples were generated from a main ranked list where the \( i \)-th component is the “true” probability \( p_i \) of belonging the instance \( i \) to the positive class. However, in contrast to [5], this list was defined considering a particular imbalance level in the assignment of true probabilities. For example, for an imbalance ratio of 1:3, the first 75% of instances in the list were linked to probabilities within the range \([0, 0.5]\) (negative class), while the rest were associated to probabilities in \((0.5, 1]\) (positive class). Given an imbalance true tuple as the one just described, a perturbed tuple was generated by randomly fluctuating the true probabilities \( p \) of negative samples within the range \([\max(0, p - \epsilon_n), \min(1, p + \epsilon_n)]\), and the true probabilities \( p \) of positive samples within the range \([\max(0, p - \epsilon_p), \min(1, p + \epsilon_p)]\). The use of two distortion terms, \( \epsilon_n \) for the negative class and \( \epsilon_p \) for the positive class, allows to simulate different scenarios of biased learning: for \( \epsilon_n > \epsilon_p \), a greater proportion of negative samples should be “misclassified”, while for \( \epsilon_n < \epsilon_p \) the positive class should be the most affected.

Table 3 is an example of a true tuple (T) with 12 samples and an imbalance ratio of 1:3, along with two derived perturbed tuples, P1 and P2, obtained from \((\epsilon_n = 0.3, \epsilon_p = 0)\) and \((\epsilon_n = 0, \epsilon_p = 0.3)\), respectively. Items typed in bold face represent misclassified samples. P1 simulates the outputs of a classifier focused on the positive class, while P2 contains the results of a biased classifier that favors the negative class.

The five collections of classifier output tuples used in the analysis were drawn from five different imbalance ratios expressed in terms of the percentage of positive samples: 5%, 10%, 15%, 20% and 25%. Each collection was composed of 130 tuples distributed in 10 per each of the 13 combinations of distortion terms ranging from \((\epsilon_n = 0.6, \epsilon_p = 0)\) to \((\epsilon_n = 0, \epsilon_p = 0.6)\) with steps \((-0.05, 0.05)\) and satisfying \(\epsilon_n + \epsilon_p = 0.6\).

<table>
<thead>
<tr>
<th>[ ]</th>
<th>0.06</th>
<th>0.11</th>
<th>0.17</th>
<th>0.22</th>
<th>0.28</th>
<th>0.33</th>
<th>0.39</th>
<th>0.44</th>
<th>0.5</th>
<th>0.67</th>
<th>0.83</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>– – – – – – – – – – – – –</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.25</td>
<td>0.52</td>
<td>0.26</td>
<td>0.49</td>
<td>0.65</td>
<td>0.24</td>
<td>0.61</td>
<td>0.67</td>
<td>0.83</td>
<td>1.0</td>
</tr>
<tr>
<td>P2</td>
<td>0.06</td>
<td>0.11</td>
<td>0.17</td>
<td>0.22</td>
<td>0.28</td>
<td>0.33</td>
<td>0.39</td>
<td>0.44</td>
<td>0.45</td>
<td>0.45</td>
<td>0.57</td>
<td>1.0</td>
</tr>
</tbody>
</table>

An independent correlation matrix between all pairs of metrics was built for each collection. Correlation coefficients were plotted in Figure 1 to make easier the understanding of results. The axes X and Y correspond to the correlation values in the range \([-1, +1]\] and the percentage of positive samples, respectively.

Several comments related with IBA_{0.05}(Gmean) can be drawn from Figure 1:

- IBA_\alpha shows a very low (negative) correlation with Acc, which has been proven not to be appropriate for imbalanced domains. Besides, the correlation coefficient of IBA_\alpha in terms of absolute value is slightly lower than those of AUC and Gmean, which are even positive.
- IBA_\alpha has a very high (positive) correlation with AUC and Gmean, suggesting that IBA_\alpha can be suitable for imbalanced distributions.
- IBA_\alpha appears to be clearly the most correlated measure with TPrate, which represents the classifier performance on the most important class (the minority one).
- IBA_\alpha presents a very low (negative) correlation with TNrate. Although AUC and Gmean show very low correlations with TNrate, their coefficients are positive.

Despite OP was defined in the context of class imbalance, it is strongly correlated with Acc, TPrate and TNrate, due to the great influence of accuracy on it.

4.2. Invariance properties

This second analysis intends to assess invariance properties of various metrics with respect to four basic changes to the confusion matrix of Table 1. A measure is said to be invariant to a certain change if it cannot distinguish a new configuration from the previous one. In general, a robust performance measure should detect every matrix transformation. Four invariance properties [9] are here used to demonstrate that IBA_\alpha is more sensitive to changes than the remaining metrics.

- **p1** invariance under the exchange of \( TP \) with \( TN \) and \( FN \) with \( FP \).
- **p2** invariance under a change in \( TN \), while all other matrix entries remain the same.
Figure 1. Scatterplot of Pearson correlation coefficients.

Table 4. Invariance properties.

<table>
<thead>
<tr>
<th>Acc</th>
<th>TPrate</th>
<th>TNrate</th>
<th>Gmean</th>
<th>AUC</th>
<th>OP</th>
<th>IBA_{\alpha}</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>p2</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>p3</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>p4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* Valid for AUC when only one classifier run is available.

5. Conclusions

We have analyzed a generalization of a new metric, IBA_{\alpha}, to evaluate the classifier performance in two-class imbalanced problems. It is defined as a trade-off between a global performance measure and a simple signed index to reflect how balanced the individual accuracies are. High values of IBA_{\alpha} are achieved when the accuracies of both classes are high and significantly balanced. Unlike most metrics, IBA_{\alpha} does not take care of the overall accuracy only, but also intends to favor classifiers with better results on the positive class.

Two theoretical studies have shown the benefits of the new metric when compared to other measures. In this sense, it has been proven that IBA_{\alpha} is strongly correlated with AUC and Gmean (generally accepted as good measures for imbalance problems). However, unlike AUC and Gmean, IBA_{\alpha} is more correlated with TPrate and less (and negatively) correlated with accuracy. Also, a study on invariance properties has shown that IBA_{\alpha} is more sensitive to changes to the confusion matrix than the other measures here considered.

References