HARD EVALUATION ANALYSIS: ACCURACY AGAINST HYBRID MEASURES FOR CLASSIFICATION TRAINING

1Hossin, M., 2Sulaiman, M.N.
1Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak (UNIMAS)
94300 Kota Samarahan, Sarawak, Malaysia
hmohamma@fit.unimas.my

2Faculty of Computer Science and Information Technology
Universiti Putra Malaysia (UPM)
43400 UPM Serdang, Selangor, Malaysia
nasir@fsktm.upm.edu.my

Abstract
As reported in literature, the use of accuracy could lead the discrimination process of optimal solution during the classification training under-performing due to its less distinctive and less discriminable value. Furthermore, the accuracy also could not perform optimally when confronted with imbalanced class problem. In contrast, the OARP and OAERP measure have shown better classification results as compared to accuracy through several comparisons and analyses. Thus, the interest of this study is to further analyze and observe under which condition one of these measures perform better than the others using hard evaluation analysis. For the evaluation purposes, this paper introduces a modified version of hard evaluation analysis. The different between the proposed evaluation method with the existing evaluation methods is, the proposed evaluation method employs a formal guideline through three performance characteristics and systematic procedure when comparing and analyzing the studied measures. In this study, three performance characteristics are value distinctiveness and discriminability, informativeness, and favors towards minority class. From the hard evaluation analysis, this study shows that both OARP and OAERP have distinctive and discriminable produced-value as compared to accuracy. However, the OAERP has shown that its produced-value much better than the accuracy and OARP in terms of informativeness, and favors towards minority class. Intuitively, the OAERP measure is a good discriminator and offer more advantages than accuracy and OARP measures in discriminating the optimal solution for balanced and imbalanced two-class classification problems.

Keywords: accuracy, hybrid measure, data classification evaluation, hard evaluation analysis

1. Introduction
The generalization ability analysis has been widely used as the baseline for the comparing measures. Typically, the generalization ability of measure is evaluated using accuracy value (overall recognition). Due to the limitation of accuracy measure, other measures such as alternative threshold measures (Japkowicz, 2008), ranking measures (Huang & Ling, 2005), hybrid measures (Caruana & Niculescu-Mizil, 2004; Ranawana & Palade, 2007), and graphical based measures (Fawcett, 2005; Hand & Till, 2001) were proposed to evaluate and compare the generalization ability of studied measures. In other studies, the results of generalization ability were also validated using statistical analysis tests (Huang & Ling, 2007; Hossin et al., 2011b) to analyze the significant difference among the studied measures.
Instead of generalization ability, many other performance characteristics could be employed as the baseline for evaluating the effectiveness of measure. In (Flach, 2003; Liu et al., 2007), the skewness analysis of measure has been proposed to evaluate the performance of studied measures. On the other hand, MacKay (2003) and Wallach (2006) have used several case studies to evaluate the quality of produced-value of studied measures subjectively in determining the best classifier. In (Huang & Ling, 2005), they proposed the statistical consistency and discriminatory analyses to evaluate and compare the advantage of produced-value of two studied measures based on degree of consistency and degree of discriminatory.

In this study, a modified version of hard evaluation method for comparing the accuracy and hybrid measures (OARP and OAERP) subjectively based on three different performance characteristics is proposed. This evaluation method adopted and adopted the evaluation method proposed by MacKay (2003) and Wallach (2006). In addition, three formal performance characteristics were defined as a guideline in conducting a better evaluation and analysis. These three performance characteristics were derived based on limitations of accuracy measure as described in the literature. The three performance characteristics are value distinctiveness and discriminability, informativeness, and favors towards minority class.

The remainder of the paper is organized as follows: Next section reviews on the accuracy and hybrid measures. The following section discusses the proposed hard evaluation method in detail. Then, this paper continues with discussion on the results analysis of the studied measures using hard evaluation method. Finally, this paper ends with conclusion and future work.

2. Literature Reviews

This subtopic starts with reviews on three studied measures which are accuracy and two hybrid measures. Then, it continues with reviews on the previous works of hard evaluation method that is used for evaluating and comparing two performance measures.

2.1 Review on studied measures

Accuracy is the most common measure that is used to build and evaluate the classifier performance in machine learning and data mining. In general, the accuracy measures the ratio of correct predictions over the total number of instances evaluated. The accuracy also known as the overall recognition rate of classifier which signifies the successfulness of classifier recognizes instances of the various classes.

Few studies have reported that the simplicity of accuracy could lead the selection and discrimination process under-performing especially when dealing with imbalanced class distribution (Chawla et al., 2004; Garcia & Herrera, 2008; Gu et al., 2009; Han, Yuan & Liu, 2009; Huang & Ling, 2005; Ranawana & Palade, 2006; Wilson, 2001; Weiss, 2004). This is because a minority class instances has less impact on the accuracy value as compared to majority class instances. The other problem of accuracy measure is its produced-value is less distinctive and less discriminable (Huang & Ling, 2005; Wallach, 2006). These shortcomings affect the discriminating power of accuracy measure in selecting and discriminating the best solution.

On the other hand, Hossin et al. (2011a, 2011b) has introduced two hybrid measures that are specific for stochastic classification training and data classification. These two measures were constructed based on combination of accuracy with the precision and recall measures. These two measures are known as an optimized accuracy with recall-precision (OARP) (Hossin et al., 2011a) and optimized accuracy with extended recall-precision (OAERP) (Hossin et al., 2011b). Empirically, both studies show that both hybrid measures were better than accuracy metric in building a better stochastic classification model and produced better predictive results. To the best of our knowledge, no work has
2.2 Review on Hard Evaluation Methods

In this study, the interest is to observe under which condition one measure performs better than the others using hard evaluation analysis. However, there were only few related studies on evaluating the performance of measures using hard evaluation analysis for data classification problems. The hard evaluation analysis was introduced by MacKay (2003). In general, hard evaluation analysis can be defined as the evaluation that compare and analyze two or more measures subjectively based on user opinions or preferences in evaluating the performance of classifiers. In (Wallach, 2006), five different performance measures (accuracy, F-measure, precision, recall and mutual information) were employed as the evaluator for comparing three classifiers. In addition, the human intuitive decision and informativeness were used as the baseline for comparing and evaluating the performance of these classifiers based on five performance measures. Based on (Wallach, 2006) analysis, it demonstrates that the mutual information measure obtained better produced-value than the other performance measures in terms of informativeness and ranked the performance of three classifiers in alignment with human intuitive decision. Meanwhile, the other performance measures could not perform like the human intuitive decision and lack of informativeness characteristic based on the given results of the three classifiers.

3. Methodology

3.1 Proposed Method

One of the weakness of MacKay (2003) and Wallach (2006) evaluation method is both methods did not utilize any formal guideline to compare the performance characteristic of studied measures. In fact, both of these evaluation methods simply rely on the human intuitive as the guideline for comparison and analysis. In addition, both of evaluation methods did not follow any systematic analysis on analyzing the performance of measures.

Due to the above weaknesses, this study proposes three formal performance characteristics as a guideline in conducting a better evaluation and analysis. These three performance characteristics were derived based on limitations of accuracy measure as described in the literature. The three performance characteristics are value distinctiveness and discriminability, informativeness, and favors towards minority class. The intuitive definitions of the three performance characteristics and rationales behind it are defined and described as below.

3.2 Formalities

For better notation, the capital letters are used to denote the total number of positive (P) and negative (N) instances in the confusion table T, whereas p and n are used for the representative number of instances (from P and N classes) covered by solution a. In addition, symbols Ψ and Ω denote balanced (P = N) and imbalanced (P < N) domain respectively. Subscript is used if more than one solution a is used. The small letter g and h is used for denoting two different measures. For simplicity and readability, the \( h(a(p,n)) \) is simplified as \( h(a) \) and omitted the argument \( (p,n) \) from solution a. For the first definition, the evaluation is applicable for both domain Ψ and Ω. However, the rest two
definitions are only applicable for domain \(Q\). This is because both situations (Definition 2 and 3) are only occurred when \(P < N\) especially when \(P \leq (5\%)\) from total \(P+N\). All of these definitions are suitable for comparing two studied measures.

3.2.1 **Definition 1 (Value Distinctiveness and Discriminability)**

For two measures \(g\) and \(h\), and two equivalent solutions \(a_1\) and \(a_2\) on domain \(\Psi\) or \(\Omega\). The value-produced by \(g\) is claimed more distinctiveness and discriminability than \(h\), if and only if, \(a_1\) and \(a_2\) evaluated by \(g\) is \(g(a_1) \neq g(a_2)\) and discriminable \((g(a_1) < g(a_2)\) or \(g(a_1) > g(a_2))\) while \(h\) is undistinguishable \(h(a_1) = h(a_2)\).

**Rationale:** Less value distinctiveness is one of the major limitations of accuracy measure. This phenomenon will cause any measure easily trapped at local optima (plateau) during the searching of an optimal solution. Thus, this limitation must be avoided by any discriminator (heuristic function). In other words, the development of future measures must be able to produce a distinctive and discriminable value for better searching and discriminating the optimal solution in huge solution space.

3.2.2 **Definition 2 (Informativeness)**

For two measures \(g\) and \(h\) and two equivalent solutions \(a_1\) and \(a_2\) where \(a_1=\{p = P, n = N-P\}\) and \(a_2=\{p = 0, n = N\}\) on domain \(\Omega\), the \(g\) is claimed more informative than \(h\), if and only if, \(g\) is able to predict \(g(a_1) > g(a_2)\) while \(h\) predicts \(h(a_1) = h(a_2)\).

**Rationale:** Another drawback of accuracy measure is there is no trade-off information between classes \((P,N)\). As a result, the accuracy measure could not differentiate the good and bad (informative and non-informative) solutions especially when two solutions are equivalent in terms of \(T(p,n)\). Based on the above definition, measure \(h\) is similar to accuracy measure. In this case, \(h\) is unable to evaluate the two solutions due to equivalent value \(h(a_1)=h(a_2)\). Although \(a_2\) able to predict all \(N\), this solution is a bad solution (non-informative) due to none \(P\) instances is covered by \(a_2\). In contrast, \(g\) values for both \(a_1\) and \(a_2\) are discriminable and able to predict and discriminate \(a_1\) and \(a_2\) based on information in both classes \((p,n)\). Since \(a_1(p) > a_2(p)\), intuitively, \(g\) is more informative than \(h\). This aspect is essential characteristics for any performance measure in discriminating the informative and optimal solution.

3.2.3 **Definition 3 (Favor towards minority class)**

For two measures \(g\) and \(h\), and three equivalent solutions \(a_1, a_2, a_3\) where \(a_1(p) > a_2(p) > a_3(p)\) on domain \(\Omega\), the \(g\) is claimed favor towards minority class than \(h\), if and only if, the value of \(g\) is favors towards minority class \((p)\) where \(g\) has the ability to predict \(g(a_1) > g(a_2) > g(a_3)\) while \(h\) undistinguishable \(h(a_1) = h(a_2) = h(a_3)\).

**Rationale:** According to Gu et al. (2009), the accuracy measure is greatly affected by the proportion of majority class \((N)\) and less impact on minority class \((P)\). Hence, it is important to employ a proper performance measure that is favors towards the minority class than the majority class. In this case, the more \(T(p)\), the better solution is produced especially for domain that is extremely imbalanced.

3.3 **Analysis Setup**

For the analysis and evaluation purposes, both hybrid measures were evaluated and compared with the accuracy measure using the definitions as described above. To ensure fair comparison, the human intuition decision was employed as a baseline for comparison discussion. All case studies employed in this analysis were manually generated and represented using \(2 \times 2\) confusion matrix. The produced-
value for both hybrid measures were calculated based on Hossin et al. (2011a) and Hossin et al. (2011b) formulas.

4. Result Analysis

4.1 Value distinctiveness and discriminability analysis

Let compare the performance of studied measures based on intuitive Definition 1. In this analysis, two case studies from two-class problem that represents balanced and imbalanced class problems were applied for comparison. For readability, the studied measures are simplified to f, g and h to signify the accuracy, OARP and OAERP measure respectively.

4.1.1 Case Study 1

Given a domain Ψ containing P=50 and N=50 instances, three measures f, g, and h, and two solutions a1 and a2. Assume that both solutions are equivalent in terms of total T(p,n) as presented in Table 1.

<table>
<thead>
<tr>
<th>s</th>
<th>tp</th>
<th>fp</th>
<th>tn</th>
<th>fn</th>
<th>tc</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>50</td>
<td>10</td>
<td>40</td>
<td>0</td>
<td>90</td>
<td>0.900000</td>
<td>0.899083</td>
<td>0.898980</td>
</tr>
<tr>
<td>a2</td>
<td>45</td>
<td>5</td>
<td>45</td>
<td>5</td>
<td>90</td>
<td>0.900000</td>
<td>0.900000</td>
<td>0.900000</td>
</tr>
</tbody>
</table>

4.1.2 Case Study 2

Given a domain Ω containing P=20 and N=80 instances, three measures f, g and h, and two solutions a1 and a2. Assume that both solutions are equivalent in terms of total T(p,n) as presented in Table 2.

<table>
<thead>
<tr>
<th>s</th>
<th>tp</th>
<th>fp</th>
<th>tn</th>
<th>fn</th>
<th>tc</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>1</td>
<td>19</td>
<td>79</td>
<td>9</td>
<td>80</td>
<td>0.800000</td>
<td>0.788538</td>
<td>0.739454</td>
</tr>
<tr>
<td>a2</td>
<td>10</td>
<td>10</td>
<td>70</td>
<td>10</td>
<td>80</td>
<td>0.800000</td>
<td>0.800000</td>
<td>0.772727</td>
</tr>
</tbody>
</table>

Based on both case studies, obviously, measure f values were not discriminable for balanced and imbalanced class problems based on two equivalent solutions. However, all hybrid measures g and h values able to produce distinctive and discriminable values as compared to f for balanced and imbalanced class problems. On the other hand, intuitively, all produced-values of hybrid measures (g and h) were better than accuracy measure (f) in terms of distinctiveness and discriminability for both case studies.

4.2 Informativeness Analysis

Let compare the performance of studied measures based on intuitive Definition 2 as baseline comparison. In this evaluation, only imbalanced domain was used for the comparison and analysis.

4.2.1 Case Study 3

Given a domain Ω containing P=5 and N=95 instances, three measures f, g and h, and two solutions a1 and a2. Assume that both solutions are equivalent in terms of total T(p,n) as presented in Table 3.
Table 3. Informativeness Analysis for Two-Class Problem using Imbalanced Class Distribution with Two Equivalent Solutions

<table>
<thead>
<tr>
<th>s</th>
<th>tp</th>
<th>fp</th>
<th>tn</th>
<th>fn</th>
<th>tc</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>0</td>
<td>5</td>
<td>95</td>
<td>0</td>
<td>95</td>
<td>0.950000</td>
<td>0.947436</td>
<td>0.850000</td>
</tr>
<tr>
<td>a2</td>
<td>5</td>
<td>0</td>
<td>90</td>
<td>5</td>
<td>95</td>
<td>0.950000</td>
<td>0.937023</td>
<td>0.934545</td>
</tr>
</tbody>
</table>

From Case Study 3, the measure $f$ was clearly not able to discriminate the best solution that most informative due to equivalent results for $a_1$ and $a_2$. Intuitively, the $a_2$ is better than $a_1$, although both solutions value were equivalent. If observed carefully, $a_1$ does not provides any useful information about the positive class ($P=0$). In contrast, $a_2$ able to provide better information for both classes. Although the values of measure $g$ was discriminable, measure $g$ predicts $a_1$ is better than $a_2$. In contrast, measure $h$ able to predict $a_2$ better than $a_1$. As conclusion, measure $h$ is better than $f$ and $g$ in terms of informativeness for two equivalent solutions.

4.2.2 Case Study 4

Let consider another case study to evaluate the informativeness characteristics of studied measures based on contradictory solutions.

Given a domain $\Omega$ containing $P=5$ and $N=95$ instances, three measures $f$, $g$ and $h$, and two solutions $a_1$ and $a_2$. Assume that both solutions are contradictory in terms of total $T(p,n)$ as presented in Table 4.

Table 4. Informativeness Analysis for Two-Class Problem using Imbalanced Class Distribution with Two Contradictory Solutions

<table>
<thead>
<tr>
<th>s</th>
<th>tp</th>
<th>fp</th>
<th>tn</th>
<th>fn</th>
<th>tc</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>0</td>
<td>0</td>
<td>95</td>
<td>5</td>
<td>95</td>
<td>0.950000</td>
<td>0.947436</td>
<td>0.850000</td>
</tr>
<tr>
<td>a2</td>
<td>5</td>
<td>6</td>
<td>89</td>
<td>0</td>
<td>94</td>
<td>0.940000</td>
<td>0.925779</td>
<td>0.922669</td>
</tr>
</tbody>
</table>

In this case study, the measure $f$ predicts that $a_1$ is better than $a_2$ based on total $T(p,n)$. Similar to Case Study 3, $a_1$ was the poor solution since no single instance from positive class is correctly predicted. Obviously, measure $f$ predicts the best solution blindly without considering the information from both classes. In addition, measure $g$ also predict $a_1$ better than $a_2$, although their produced-values were discriminable. On the other hand, measure $h$ able to predict $a_2$ better than $a_1$, even though the total of $a_2(p,n)$ was lower than $a_1(p,n)$.

Based on both case studies, this study can conclude that OAERP measure is better than accuracy and OARP in terms of informativeness using two equivalent and contradictory solutions and also predicts similar to human intuition.

4.3 Favors towards minority class

Let consider another case study for evaluating the intuitive Definition 3. For this analysis, two extremely imbalanced class distributions were proposed for the evaluation ($(P=5, N=95)$ and $(P=5, N=9995)$).

4.3.1 Case Study 5

Given a domain $\Omega$ containing $P=5$ and $N=95$ instances, three measures $f$, $g$ and $h$, and six solutions $a_1$, $a_3$, $a_4$, $a_5$ and $a_6$. Assume that all solutions are equivalent in terms of total $T(p,n)$ as presented in Table 5.
Table 5. Favors towards Minority Class Analysis for Two-class Problem using Imbalanced Class Problem (5:95)

<table>
<thead>
<tr>
<th>s</th>
<th>tp</th>
<th>fp</th>
<th>tn</th>
<th>fn</th>
<th>tc</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>a₁</td>
<td>0</td>
<td>0</td>
<td>95</td>
<td>5</td>
<td>95</td>
<td>0.950000</td>
<td>0.947436</td>
<td>0.850000</td>
</tr>
<tr>
<td>a₂</td>
<td>1</td>
<td>1</td>
<td>94</td>
<td>4</td>
<td>95</td>
<td>0.950000</td>
<td>0.939817</td>
<td>0.900822</td>
</tr>
<tr>
<td>a₃</td>
<td>2</td>
<td>2</td>
<td>93</td>
<td>3</td>
<td>95</td>
<td>0.950000</td>
<td>0.946846</td>
<td>0.913032</td>
</tr>
<tr>
<td>a₄</td>
<td>3</td>
<td>3</td>
<td>92</td>
<td>2</td>
<td>95</td>
<td>0.950000</td>
<td>0.947056</td>
<td>0.922056</td>
</tr>
<tr>
<td>a₅</td>
<td>4</td>
<td>4</td>
<td>91</td>
<td>1</td>
<td>95</td>
<td>0.950000</td>
<td>0.941723</td>
<td>0.929010</td>
</tr>
<tr>
<td>a₆</td>
<td>5</td>
<td>5</td>
<td>90</td>
<td>0</td>
<td>95</td>
<td>0.950000</td>
<td>0.937023</td>
<td>0.934545</td>
</tr>
</tbody>
</table>

4.3.2 Case Study 6

Given a domain Ω containing P=5 and N=9995 instances, three measures f, g and h, and six solutions a₁, a₂, a₃, a₄, a₅, and a₆. Assume that all solutions are equivalent in terms of total T(p,n) as given in Table 6.

Table 6. Favors towards Minority Class Analysis for Multi-class Problem using Extremely Imbalanced Class Problem (5:9995)

<table>
<thead>
<tr>
<th>s</th>
<th>tp</th>
<th>fp</th>
<th>tn</th>
<th>fn</th>
<th>tc</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>a₁</td>
<td>0</td>
<td>0</td>
<td>9995</td>
<td>5</td>
<td>95</td>
<td>0.999500</td>
<td>0.999475</td>
<td>0.899500</td>
</tr>
<tr>
<td>a₂</td>
<td>1</td>
<td>1</td>
<td>9994</td>
<td>4</td>
<td>95</td>
<td>0.999500</td>
<td>0.988398</td>
<td>0.949508</td>
</tr>
<tr>
<td>a₃</td>
<td>2</td>
<td>2</td>
<td>9993</td>
<td>3</td>
<td>95</td>
<td>0.999500</td>
<td>0.996055</td>
<td>0.961415</td>
</tr>
<tr>
<td>a₄</td>
<td>3</td>
<td>3</td>
<td>9992</td>
<td>2</td>
<td>95</td>
<td>0.999500</td>
<td>0.996277</td>
<td>0.970345</td>
</tr>
<tr>
<td>a₅</td>
<td>4</td>
<td>4</td>
<td>9991</td>
<td>1</td>
<td>95</td>
<td>0.999500</td>
<td>0.990417</td>
<td>0.977289</td>
</tr>
<tr>
<td>a₆</td>
<td>5</td>
<td>5</td>
<td>9990</td>
<td>0</td>
<td>95</td>
<td>0.999500</td>
<td>0.985227</td>
<td>0.982844</td>
</tr>
</tbody>
</table>

From both case studies, intuitively, a₆ was better than five other solutions. Based on intuition and given intuitive Definition 3, all solutions from both case studies could be ranked as follows: \(a₆>a₅>a₄>a₃>a₂>a₁\). In this case, all solutions were ranked decreasingly based on their performance on minority class p. The more \(T(p)\) the solution obtained, the better solution towards minority class. For both case studies, the measure f could not distinguish which solution was the best. If compared to measures g and h, all solutions values were discriminable. However, the produced-values of measure g was not aligned with intuition ranked (ill-ranked). Measure g ranked the best solution at the last positioned. Similar to the earlier informativeness analysis (Table 4), the measures g ranked the poor solution \(a₁\) at first positioned. For measure g, its solutions were ranked as follows: \(g = (a₆>a₅>a₄>a₃>a₂>a₁)\). In contrast, for measure h, all solutions were ranked similar to intuition ranked (well-ranked). Clearly, both case studies indicate that OAERP measure is more favors towards the minority class than accuracy, and OARP measures.

4.4 Result Summary

Based on all discussed case studies, it shows that the produced-values of both hybrid measures are distinctive and discriminable as compared to accuracy value. However, only OAERP measure able to demonstrate better performance on the informativeness and favors towards minority class. Obviously, the OAERP measure is a good discriminator and offer more advantages than accuracy and OARP measure in discriminating the best solution for two-class problem including balanced and imbalanced class problems. In addition, this study also shows that the accuracy measure is the powerless measure since all solutions were not discriminable. For OARP measure, the produced-values for all solutions

Organized by http://WorldConferences.net
were discriminable but less informative and less favors towards the minority class (ill-ranked). However, the OARP measure is competitively better than accuracy measure in some aspects as explained above.

5. Conclusion and Future Work

In this paper, the modified version of hard evaluation method was proposed successfully to evaluate the studied measures subjectively based three different performance characteristics. Interestingly, no classification algorithm was used for comparing the effectiveness of the studied measures in discriminating the best solution. From the analysis, this study concludes that both hybrid measures have distinctive and discriminable produced-value as compared to the accuracy measure. However, the OAERP measure has shown that its produced-value is better than the accuracy and OARP in terms of informativeness, and favors towards minority class. However, the consistency and discriminating power of each studied measure was not evaluated in this evaluation analysis. In reality, the consistency and discriminating power of each studied measure are difficult to evaluate subjectively since the definition of hard (strict) discriminating power is restricted by many factors. Therefore, for future work, this study attempts to identify a suitable evaluation method for measuring the effectiveness of studied measures in terms of consistency and discriminating power.

References


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


