Machine Learning and Data Mining in Pattern Recognition
Volume 7988 of the series Lecture Notes in Computer Science pp 281-295. The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-642-39712-7_22.

Pacc - A Discriminative and Accuracy Correlated Measure for Assessment of Classification Results

Madhav Sigdel and Ramazan Aygun

University of Alabama in Huntsville, Department of Computer Science, Huntsville, AL, USA 35899

Abstract. Measuring the performance of a classifier properly is important to determine which classifier to use for an application domain. Many performance (correctness) measures have been described to facilitate the comparison of classification results. However, the comparison is not straightforward since different experiments may use different datasets, different class categories, and different data distribution, thus biasing the results. In this paper, we provide an overview of the widely used classifier performance measures and list the qualities expected in a good performance measure. We introduce a novel measure, *Probabilistic accuracy (Pacc)*, to compare multi-class classification results and make a comparative study of several measures and our proposed method based on different confusion matrices. Experimental results show that our proposed method is discriminative and highly correlated with accuracy compared to other measures. The web version of the software is available at http://sprite.cs.uah.edu/perf/.

1 Introduction

There are a number of factors that affect the performance of a classification problem: the classification algorithm, features, the number of classes, the datasets, and the data distribution. The performance of a classification methodology should be compared and analyzed to select a particular method based on usefulness of the classifier. Result of a classification problem is often represented in the form of a matrix called confusion matrix. Thus, the performance of two classifiers can be evaluated by comparing the corresponding confusion matrices.

The most common method for the evaluation of classification results is to compute a performance metric based on the confusion matrix. Several such measures have been described in the literature. Most of these measures have been developed for binary classification. Accuracy is one of the widely used measures which is the percentage of correct decisions made by the classifier. However the overall accuracy is not a very reliable measure for problems such as protein crystallization classification [1] where the cost of misclassifying crystals as non-crystals is very high or the proportion of data in the different categories is significantly different. There are also other measures like sensitivity, specificity, precision and F-measure are formulated for binary classification. Some

research [2], [3] describes methods to extend F-measure for multi-class classification. Research studies [4], [5], [6] proposes extensions to area under the ROC curve (AUC) for multiclass result evaluation. There are also other measures like confusion entropy [7] and K-category correlation coefficient [8] that are naturally applicable to the performance evaluation of multiclass classification results.

Analysis based on multiple performance measures is also another popular method for evaluation of classifiers. For example, percision and recall are often analyzed together. For using multiple measures, a problem is that when comparing classifier A and classifier B, classifier A may outperform classifier B with respect to one measure, while classifier B may outperform classifier A with the other measure.

The advantages and disadvantages of the widely used performance measures like accuracy, precision, recall, correlation coefficient, relative entropy, etc. are analyzed in [9] and [10]. Sokolova et al. mention that different performance measures possess invariance properties with respect to the change in a confusion matrix and these properties can be beneficial or adverse depending on the problem domain and objectives [9]. Statistical techniques for comparison of classifiers over multiple datasets are described in [11] and [12]. Perner [13] describe a methodology for interpreting results from decision trees. Though there are research on measures for classification results, a comparative study of these measures with classification results for binary and multi-class classification have not been explored much.

In this paper, we attempt to analyze classification performance measures based on a number of classification results. It should be noted that the performance here is related to the correctness of the classifier and not in terms of speed or efficiency. We try to analyze the consistency between different measures and also the degree of discrimination for confusion matrix comparison. We propose a new measure, *probabilistic accuracy (Pacc)*, based on the difference in probability of correct classification and probability of misclassification given a confusion matrix. Accuracy measure is still one of the widely used measures despite its limitations. A major reason for this is that the accuracy measure has simple semantic correspondence to our understanding. For some other measures, a high (or sometimes low) value is preferred and it is hard to derive a semantic meaning from those measures. Therefore, we develop our measure in a way that it is consistent with accuracy but more discriminative than accuracy. Besides it is defined for every type of confusion matrix (binary or multi-class) with valid values and is less susceptible to scaling of the number of items in a class. The web-version of the software is available online at http://sprite.cs.uah.edu/perf/ which allows computation of *Pacc* measure along with other popular performance measures.

The rest of this paper is organized as follows. Section 2 provides an overview of the classification performance measures for binary classification and multi-class classification. Section 3 discusses about the qualities expected in a good performance measure for classification results comparison. Section 4 provides the formal definition and semantics of Pacc measure. Section 5 provides a compar-

ative study of several performance measures and our proposed measure considering several cases of confusion matrices. Section 6 concludes the paper.

2 Performance Measures Overview

Broadly there are three methods for comparing two confusion matrices. The first method is to compare corresponding elements of the two matrices. Confusion matrices may be normalized so that individual elements become comparable. The second method involves computing a function f that takes confusion matrix as the input and returns a single metric. The comparison of two matrices $M1$ and $M2$ then involves computation of $f(M1)$ and $f(M2)$. Depending on the measure, a high or low value can represent a good or bad classification result. The third approach is to compute several measures and analyze the results together. For example, precision and recall are often analyzed together. For using multiple measures, a problem is that when comparing classifier A and classifier B, classifier A may outperform classifier B with respect to one measure, while classifier B may outperform classifier A with the other measure.

In this paper, we focus on methods that allow analysis based on a single value. Such measures can be grouped as measures for binary classification and measures for multi-class classification. Note that the measures for multi-class classification are also applicable for binary classification. For binary classification, accuracy, sensitivity (also called as recall or hit rate), specificity, precision, F-measure, and Kappa statistic are used in practice. Accuracy is the percentage of correct decisions made by a classifier. Sensitivity is the ratio of correctly predicted positives to the actual positives. This is also called recall or hit rate. Specificity is the ratio of correctly predicted negatives to the actual negatives. Precision is the ratio of correctly predicted positives to the total positives. F-measure is defined as the harmonic mean of precision and recall for binary classification. Kappa statistic is defined as the proportion of agreement between two rankings corrected for chance [14].

The result of an N-class classification experiment with classes 0..N-1 can be visualized in matrix of size N x N. This matrix is called confusion matrix, contingency matrix, or contingency table. Matrix C represents a generalized N x N confusion matrix. The value C_{ij} refer to the number of items of i^{th} class classified as j^{th} class where i represents the actual class and j represents the predicted class. It is easy to see that the elements in the diagonals represent the number of items of each class correctly classified. Thus, the ideal case would be a diagonal matrix (i.e., each cell except the diagonals are equal to zero) meaning a perfect classification.

$$C = \begin{pmatrix} C_{00} & C_{01} & \ddots & C_{0(N-1)} \\ C_{10} & C_{11} & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots \\ C_{(N-1)0} & C_{(N-1)1} & \ddots & C_{(N-1)(N-1)} \end{pmatrix}$$

2.1 Confusion Entropy

Wei et al. introduce confusion entropy method as a performance measure for multi-class classification [7]. The authors apply the concept of probability and information theory for the calculation of confusion entropy. The misclassication probability of classifying samples of class i to class j subject to class j is denoted by P_{ij}^{j} and is given by (1). Similarly, misclassication probability of classifying samples of class i to class j subject to class i is denoted by P_{ij}^{i} and is defined as in (2).

$$P_{ij}^{j} = \frac{C_{ij}}{\sum_{k=0}^{N-1} C_{jk} + C_{kj}} i \neq j, i, j = 0..N-1 \tag{1}$$

$$P_{ij}^{i} = \frac{C_{ij}}{\sum_{k=0}^{N-1} C_{ik} + C_{ki}} i \neq j, i, j = 0..N-1 \tag{2}$$

Confusion entropy of class j is defined as in (3).

$$CEN_j = -\sum_{k=0, k\neq j}^{N-1} P_{jk}^{j} log_{2(N-1)} P_{jk}^{j} + P_{kj}^{j} log_{2(N-1)} P_{kj}^{j} \tag{3}$$

Overall entropy (CEN) is defined by (4). Note that P_{ii}^{i}=0.

$$CEN = \sum_{j=0}^{N-1} P_j CEN_j \tag{4}$$

where P_j is defined as in (5):

$$P_j = \frac{\sum_{k=0}^{N-1} C_{jk} + C_{kj}}{2 \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} C_{kl}} \tag{5}$$

The value of CEN ranges from 0 to 1 with 0 signifying the best classification and 1 indicating the worst classification.

2.2 K-Category Correlation Coefficient

Gorodkin proposes K-category correlation coefficient to compare two confusion matrices [8]. The method utilizes the concept of covariance and tries to compute the covariance between actual K-category assignment and the observed assignment. Consider two matrices X, Y of size N x K where N is the number of items and K is the number of categories. Let matrix X and matrix Y represent

the actual assignment and predicted assignment, respectively. The correlation coefficient R_k is defined as (6).

$$R_k = \frac{cov(X,Y)}{\sqrt{cov(X,X)}\sqrt{cov(Y,Y)}} \tag{6}$$

In terms of confusion matrix as denoted in the beginning of this section, the covariances can be written as follows:

$$cov(X,Y) = \sum_{k,l,m=0}^{N-1} C_{kk}C_{ml} - C_{lk}C_{km} \tag{7}$$

$$cov(X,X) = \sqrt{\sum_{k=0}^{N-1}\left(\sum_{l=0}^{N-1} C_{lk}\right)\left(\sum_{f,g=0,f\neq k}^{N-1} C_{gf}\right)} \tag{8}$$

$$cov(Y,Y) = \sqrt{\sum_{k=0}^{N-1}\left(\sum_{l=0}^{N-1} C_{kl}\right)\left(\sum_{f,g=0,f\neq k}^{N-1} C_{fg}\right)} \tag{9}$$

The value of R_k ranges from -1 to +1 with +1 indicating the best classification and -1 indicating the worst classification.

2.3 F-measure for multiclass problems

F-measure combines the two metrics - recall and precision and is defined as the harmonic mean of the two. As described in [3], the recall (R_i), precision (P_i), and F-measure (F_i) for class i in a multiclass problem can be defined by following equations:

$$P_i = \frac{TP_i}{TP_i + FP_i}, \quad R_i = \frac{TP_i}{TP_i + FN_i}, \tag{10}$$

$$(F_i) = \frac{2P_iR_i}{P_i + R_i} \tag{11}$$

where TP_i is the number of objects from class i assigned correctly to class i, FP_i is the number of objects that do not belong to class i but are assigned to class i, and FN_i is the number of objects from class i predicted to another class. To compute the overall F-measure, macro-averaging and micro-averaging are used. Macro-averaged F-measure, F(macro), is calculated as the average of F-measure for each category. Micro-averaged F-measure, F(micro), aggregates the recall and precision of classes.

$$F(macro) = \frac{1}{N}\sum_{i=0}^{N-1} F_i, \quad F(micro) = \frac{2PR}{P+R} \tag{12}$$

where P and R are defined by the following equations:

$$P = \frac{\sum_{i=0}^{N-1} TP_i}{\sum_{i=0}^{N-1} TP_i + FP_i}, \quad R = \frac{\sum_{i=0}^{N-1} TP_i}{\sum_{i=0}^{N-1} TP_i + FN_i} \tag{13}$$

2.4 Kappa statistic

Kappa statistic is defined as the proportion of agreement between two rankings corrected for chance [14]. In the context of classification result, the agreement between the actual categories and predicted categories forms the basis for calculation of Kappa. Let $S = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} C_{ij}$ represent the total number of items in the confusion matrix, $C_{i.} = \sum_{j=0}^{N-1} C_{ij}$ represent the i^{th} row marginal and $C_{.i} = \sum_{j=0}^{N-1} C_{ji}$ represent the i^{th} column marginal. Then, Cohen's Kappa (K) is given by (14).

$$K = \frac{P_o - P_e}{1 - P_e} \tag{14}$$

where $P_o = \frac{1}{S} \sum_{i=0}^{N-1} C_{ii}$ is the proportion of agreement between observed and actual categories, and $P_e = \frac{1}{S^2} \sum_{i=0}^{N-1} C_{i.}C_{.i}$ is the proportion of observations for which agreement is expected by chance. $P_o - P_e$ is the proportion of agreement beyond what is expected by chance, and $1 - P_e$ is the maximum possible proportion of agreement beyond what is expected by chance. Values of Kappa can range from -1 to +1, with -1 indicating perfect disagreement below chance, and +1 indicating perfect agreement above chance.

3 Qualities of good performance measure

Consider the following confusion matrices.

$$M = \begin{pmatrix} 70 & 10 \\ 10 & 10 \end{pmatrix} \quad N = \begin{pmatrix} 80 & 0 \\ 20 & 0 \end{pmatrix}$$

Matrix M has 10 items of each class misclassified and Matrix N has all items of class 0 correctly classified while none of the items in class 1 are correctly classified. The accuracy for both matrices is 80%, however the classifier that results N might not be useful since all items have been classified to a single class. This is because the distribution of misclassification is ignored by the measure.

Consider other hypothetical classification results given by matrix K and L.

$$K = \begin{pmatrix} 20 & 0 \\ 20 & 10 \end{pmatrix} \quad L = \begin{pmatrix} 100 & 0 \\ 20 & 10 \end{pmatrix}$$

The first category has all items correctly classified while 20 of 30 objects are misclassified for the second category. This could be the case where it is easy to classify the objects of the first category. Suppose the data items of the first category are increased by 5 folds and the new confusion matrix is given by matrix L.

The accuracy of the experiment is increased from 60% to 84% just by increasing the items in the first category which can be misleading. We expect the performance measure to be less susceptible to scaling of the dataset.

In this section, we list the qualities expected in a good performance measure for the evaluation of classification results. Some of these qualities may or may not be desired depending on the application. Therefore, we first list the desired qualities in a good performance metric irrespective of the problem domain. These are listed as follows:

- The measure should have the highest value for the best case i.e., when all items correctly classified. There can be many varieties of the the worst case depending on the distribution of misclassification. We may want to distinguish those cases.
- The measure should not be affected by a scale factor. This means that the measure for matrix C should be same as the measure for a x C where a is a scale factor.
- The measure should be based on all the values in the confusion matrix for the calculation.
- The measure should be able to distinguish different confusion matrices. The value should decrease with increase in misclassified cases and increase with decrease in misclassified cases or vice versa.
- It should be useful for classification with any number of classes.

Likewise, there can be other desired qualities depending on the application. These are listed as follows.

- If the important class occurs very rarely, performance measures are affected by the scaling of data. Thus we desire that a performance measure should be less affected by the scaling of one or more of the classes as long as the distribution of misclassification is proportional. However, in some cases, we may want to pay extra attention to a class which is more likely than others.
- Misclassification into a single class may be considered better than misclassification into several classes. If the misclassification occurs into a single class, the classifier may be tuned by focusing on the problematic classes.

4 Proposed Pacc measure

We introduce *Probabilistic accuracy (Pacc)* measure for comparison of two N-class confusion matrices. This measure is based on the difference in the probability of correct classification and the probability of misclassification. C_{ij} refers to the number of items in class i that are classified to class j. The occurrence (probability) of C_{ij} is related to both the number of items in class i and the number of items of other classes that are classified to class j. P_{ij} is the probability of occurrence of C_{ij} subject to actual class i and observed class j and is defined as in (15). Apparently, C_{ij} is contained in both sub-parts of the denominator. If $i \neq j$, P_{ij} is the probability of misclassifying item of class i subject to class j. P_{ij} should increase if the majority of incorrect classifications into class j are coming from items in class i. Note that the numerator does not contain the correctly classified cases. Likewise, the probability of correctly classifying items, denoted by P_{ii}, can be defined as (16). Here the numerator consists only the correctly classified cases i.e., the diagonal elements of the confusion matrix.

$$P_{ij} = \frac{2C_{ij}}{\sum_{k=0}^{N-1} C_{ik} + C_{kj}} i \neq j, k = 0, ..N-1 \tag{15}$$

$$P_{ii} = \frac{2C_{ii}}{\sum_{k=0}^{N-1} C_{ik} + C_{ki}} \tag{16}$$

Maximum value for any P_{ij} is 1. This occurs when all items of class i are classified solely to class j and none of the items from other classes are classified to class j. In other words, this is like pure misclassification probability between class i and j which indicates that all items of class i are classified as class j and all observed class j classifications result from class i items. The minimum value is 0 which occurs when all items in a class are correctly classified and no items of other classes is predicted to this class.

Now we define two terms: error (ϵ) and correctness (c) in terms of P_{ii} and P_{ij} as given in (18) and (17), respectively. Error probability (ϵ) is the average of the probabilities of misclassification and correctness probability (c) is the average of the probabilities for correct classification. c and ϵ lie between 0 and 1. High correctness probability and low error probability are desired for good classification. The difference between c and ϵ yields a value between -1 to +1. It is normalized to the range 0 to 1 as in (19) so that it can be correlated with the accuracy measure.

$$\epsilon = \frac{1}{N} \sum_{i=0}^{N-1} \sum_{j=0, i \neq j}^{N-1} P_{ij} \tag{17}$$

$$c = \frac{1}{N} \sum_{i=0}^{N-1} P_{ii} \tag{18}$$

Table 1. Classification results for 2-class problem

	O_0	O_1	O_0	O_1	O_0	O_1	O_0	O_1	O_0	O_1
	A		B		C		D		E	
C_0	50	0	25	25	50	0	10	40	0	50
C_1	0	50	25	25	50	0	40	10	50	0
	F		G		H		I		J	
C_0	80	0	70	10	80	0	40	40	0	80
C_1	0	20	10	10	20	0	10	10	20	0

Table 2. Performance measures for matrices A to J in 1

MATRIX	ACC	KAPPA	R_k	1-CEN	FMEAS	PACC
A	1.00	1.00	1.00	1.00	1.00	1.00
B	0.50	0.00	0.00	0.00	0.50	0.50
C	0.50	0.00	NAN	0.60	NAN	0.50
D	0.20	-0.60	-0.60	-0.06	0.20	0.20
E	0.00	-1.00	-1.00	0.00	NAN	0.00
F	1.00	1.00	1.00	1.00	1.00	1.00
G	0.80	0.37	0.38	0.40	0.69	0.74
H	0.80	0.00	NAN	0.68	NAN	0.64
I	0.50	0.00	0.00	0.17	0.45	0.50
J	0.00	-0.47	-1.00	0.28	NAN	0.00

$$Pacc = \frac{1}{2} + \frac{c - \epsilon}{2} \tag{19}$$

The value of Pacc is maximum (i.e., 1) when all the items are correctly classified. In this case, ϵ is 0 because every C_{ij} where i $\neq$ j is 0. Likewise, c is equal to 1. Hence, the difference between c and ϵ is 1 and the normalization to [0-1] gives the value 1.

$$\epsilon = \frac{1}{N} \sum_{i=0}^{N-1} \sum_{j=0, i \neq j}^{N-1} P_{ij} = 0$$

$$c = \frac{1}{N} \sum_{i=0}^{N-1} P_{ii} = \frac{1}{N} \sum_{i=0}^{N-1} \frac{2C_{ii}}{C_{ii} + C_{ii}} = \frac{1}{N} \sum_{i=0}^{N-1} 1 = 1$$

$$Pacc = \frac{1}{2} + \frac{c - \epsilon}{2} = \frac{1}{2} + \frac{1 - 0}{2} = 1$$

The value of Pacc is minimum (i.e., 0) when every items from a class are misclassified to a unique single class.

5 A Comparative Study of Performance Measures

In this section, we perform a comparative study of the following performance measures: accuracy (ACC), Kappa statistic (KAPPA), K-category correlation coefficient (R_k), confusion entropy (CEN), macro-averaged F-measure (FMEAS) and our Pacc measure for several confusion matrices. The measure for CEN is subtracted from 1 to simplify the comparison (for this measure the lowest value (i.e., 0) is the best and the highest value (i.e., 1) is the worst).

5.1 Analysis of Measures for 2-class Classification

Consider the classification results for the 2-class problems as given by confusion matrices A to J in Table 1. The matrices follow the generalized confusion matrix structure outlined in Section 2. The columns O_i indicate the objects classified to class i and the rows C_i indicate the actual categories. 2 shows the results of these measures.

Accuracy does not account for the distribution of misclassified items. As long as the numbers of correct predictions remain the same, accuracy remains the same. In matrix G, the accuracy is 80% where half of the items in class 1 are misclassified to class 0. Similarly, in matrix H, the accuracy is still 80% where all items of one class have been classified to other class. Thus analysis based on accuracy measure can be misleading.

CEN values are proposed to be in the range [0..1]. However, for some cases in binary classification we get values of CEN that cannot be interpreted. For matrix B in 2, the value of (1-CEN) is 0 signifying the worst classification. However, Matrix B has half the items in each class correctly classified. Moreover, the value of CEN goes out of range in some cases. For matrix D, CEN value is 1.06 which is out of the range. Such cases arise when the ratio of correct cases to incorrect cases is less than 1 for both the categories. Matrix C looks good since it is close to 0; however, this matrix has all items classified to a single class which may not be useful. Among the matrices G, H, I and J, we would expect the measure to indicate G as the best classification and J as the worst classification. The CEN values indicate matrix H to be the best among the four and matrix I to be the worst. Definitely, the performance measure for J should have the worst value as it has all the items misclassified. This shows that CEN is not a reliable measure.

The value NaN in the R_k columns corresponding to matrix C and matrix H indicates that it is not a number (NaN). As NaN can be obtained in many cases, it becomes difficult to compare the results.

Similarly, there are several cases where the F-measure is undefined. Such cases arise when there are some categories for which no correct classification is made. Moreover, we can observe that NaN does not necessarily occur when the confusion matrix has low accuracy. As can be seen the value of FMEAS is NaN for matrix H and matrix J. The corresponding accuracy for H and J are 0.80 and 0, respectively.

Table 3. Classification results for 3-class problem

	O_0	O_1	O_2	O_0	O_1	O_2	O_0	O_1	O_2	O_0	O_1	O_2
	A			B			C			D		
C_0	60	0	0	40	0	20	30	30	0	30	15	15
C_1	0	60	0	0	60	0	0	60	0	0	60	0
C_2	0	0	60	0	0	60	0	0	60	0	0	60
	E			F			G			H		
C_0	40	10	10	0	30	30	20	20	20	0	0	60
C_1	10	40	10	0	60	0	20	20	20	0	60	0
C_2	10	10	40	0	0	60	20	20	20	60	0	0

Table 4. Performance measures for matrices A to H in 3

MATRIX	ACC	KAPPA	R_k	1-CEN	FMEAS	PACC
A	1.00	1.00	1.00	1.00	1.00	1.00
B	0.89	0.83	0.85	0.86	0.89	0.90
C	0.83	0.75	0.78	0.84	0.82	0.84
D	0.83	0.75	0.78	0.76	0.82	0.83
E	0.67	0.50	0.50	0.40	0.67	0.67
F	0.67	0.50	0.58	0.72	NAN	0.63
G	0.33	0.00	0.00	0.14	0.33	0.33
H	0.33	0.00	0.00	0.67	NAN	0.33

For the binary cases with balanced distribution, our Pacc measure is consistent with the accuracy. For the unbalanced cases, our Pacc provides different values and is more discriminative than other measures.

5.2 Analysis of measures for 3-class classification

Consider 3-class classification results (confusion matrices A to H) in 3 with balanced distribution of items in each class. The performance measures for these matrices are provided in 4. Figure 1 shows the plot of these measures for the corresponding matrices. As we go from matrix A to matrix H, the number of misclassified items is increased. Therefore, we expect similar changes in the performance measure.

From the performance measures in 4 and graph in Fig. 1, we observe the following.

- Accuracy measure is not discriminative. For example, matrices C and D, E and F, and G and H have the same accuracy. Therefore, we cannot distinguish among those just based on accuracy.
- Confusion entropy measure is not consistent with other measures. Confusion entropy measure suggests confusion matrices F and H being better than E which does not look correct. Likewise, the results are not as expected when

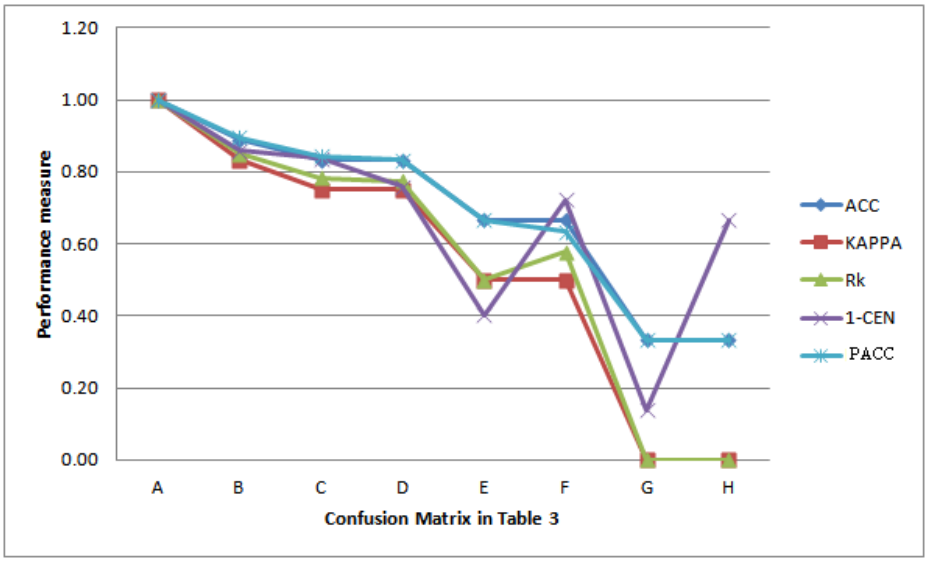

Fig. 1. Graph showing the plot of performance measures from 4. Column MATRIX in the table 4 correspond to the confusion matrices in 3

.

we compare matrix E and H. Matrix E has 40 items in each category correctly classified. On the other hand, Matrix H has none of the items in the first category and third category correctly classified. CEN suggests matrix H to be better result compared to matrix E. Therefore, the result is not as desired. Another observation is that CEN considers misclassification to a single class to be better than misclassification to several classes. This property may or may not be desired depending on the application.

- F-measure is not computable for the confusion matrices F and H and is less discriminative compared to Pacc measure.
- The measures R_k, F-Measure, and Pacc are discriminantive than accuracy. Pacc follows the decreasing trend of values as we go from error matrix A to H.

5.3 Comparative evaluation of performance measures

Discriminative property: One important requirement for a performance measure is the ability to distinguish confusion matrices. From the examples presented in earlier sections, it is seen that measures like accuracy and Kappa statistic have low discriminative power. To analyze the discriminative power of the various measures, we considered a 3-class problem with 5 items in each category. 21 combinations are possible for the distribution of 5 items into different categories. Therefore, a total of 9261 (21 x 21 x 21) confusion matrices are possible.

Table 5. Table showing the count of distinct values from 9261 possible confusion matrices in a 3-class problem with 5 items in each category

Measure	Num distinct values	Avg(abs diff with acc)
ACC	16	-
KAPPA	16	0.166
R_k	183	0.166
CEN	1504	0.359
FMEAS	368	0.169
Pacc	669	0.029

We calculated all the measures for these confusion matrices. 5 shows the count of distinct values obtained for 9261 confusion matrices. Accuracy and Kappa statistic have the lowest discrimiative power as both of these measures have only 16 possible values for all of these matrices. Confusion entropy has the largest number of distinct values. Pacc measure also has high discriminative power.

NaN values and our resolution: Measures like F-measure and correlation coefficient may produce NaN as the result. For the 9261 possible confusion matrices in a 3-class problem with 5 items in each category, F-measure is NaN for almost 63% of the confusion matrices. If both precision and recall are 0, F-measure becomes undefined or NaN. Therefore, when these measures are used for the evaluation of classification results, necessary patches should be applied so that NaN is not an output. One approach to solve this would be to take the measure to be equal to 0. Nonetheless, there are multiple cases for the measure to be 0 making it difficult to distinguish/rank classification results. Correlation coefficient can produce NaN in case all the items are classified to a single class. If all items are classified into a single class, the variance for that class is 0. Since there are no items that are classified into other classes, the variance for those classes are also 0. This corresponds to a column in confusion matrix with non-zero values where the rest of the values are 0 in the confusion matrix.

Accuracy correlation: 5 provides the average of absolute difference between accuracy and other measure for the 9261 confusion matrices. For the correlation coefficient and Kappa statistic, the final average value is divided by 2 since its original range is from -1 to +1. The difference is the least for Pacc measure thus revealing a high correlation of Pacc measure with accuracy . The inconsistency in confusion entropy measure is also reflected by the low correlation with accuracy. Kappa, R_k, and F-measure also have the lower correlation with accuracy compared to Pacc measure.

Scale invariance: A new set of confusion matrices were created by scaling the confusion matrices A to H provided in 3. For each of these matrices, the second row is doubled and the third row is increased by 5 times. The performance measures for the modified matrices are presented in 6. Figure 2 shows the plot of the difference in the performance measures in 4 and 6 (i.e., the difference in the measures between original and scaled matrices). F-measure is not included in this plot as there are some values for F-measure that are undefined. From the

Table 6. Performance measures for matrices A to H in 3 with second row increased twice and 3rd row increased by 5 times

MATRIX	ACC	KAPPA	R_k	1-CEN	FMEAS	Pacc
A	1.00	1.00	1.00	1.00	1.00	1.00
B	0.96	0.92	0.92	0.92	0.92	0.93
C	0.94	0.88	0.89	0.93	0.85	0.88
D	0.94	0.88	0.88	0.89	0.86	0.89
E	0.67	0.44	0.46	0.46	0.61	0.65
F	0.88	0.75	0.77	0.85	NaN	0.73
G	0.33	0.00	0.00	0.23	0.30	0.35
H	0.25	0.04	0.06	0.76	NaN	0.33

Table 7. Table listing the properties of various measures

Measure	Discriminative	NaN values	Accuracy correlation	Scale invariance
ACC	Low	No	-	Low
KAPPA	Low	No	Low	Low
R_k	Medium	Yes	Low	Medium
CEN	High	No	Very Low	High
FMEAS	Medium	Yes	Low	High
Pacc	High	No	High	High

figure, we can see that accuracy and K-category correlation coefficient (R_k) are the most affected measures by the scaling. CEN measure and Pacc measure are comparatively less affected. However, CEN has other inconsistency problems as outlined earlier.

7 provides a summary of various properties exhibited by the different measures. The level of discrimination and the level of scale invariance for accuracy is considered to be low. The levels for other measures are assigned relative to the accuracy. Pacc measure compares best among others as it has high level of discriminancy, does not result NaN values, scale invariance is high, and it is highly correlated with the accuracy measure.

6 Conclusion

In this paper, we explained the difficulties in comparing two classification experiments and highlighted the need for a good performance measure. We listed expected qualities in a good classifier performance measure and introduced a novel measure Probabilistic accuracy (Pacc) which is based on the difference between probabilities of correct and incorrect classification. We made a comparative analysis of widely used performance measures and our proposed method

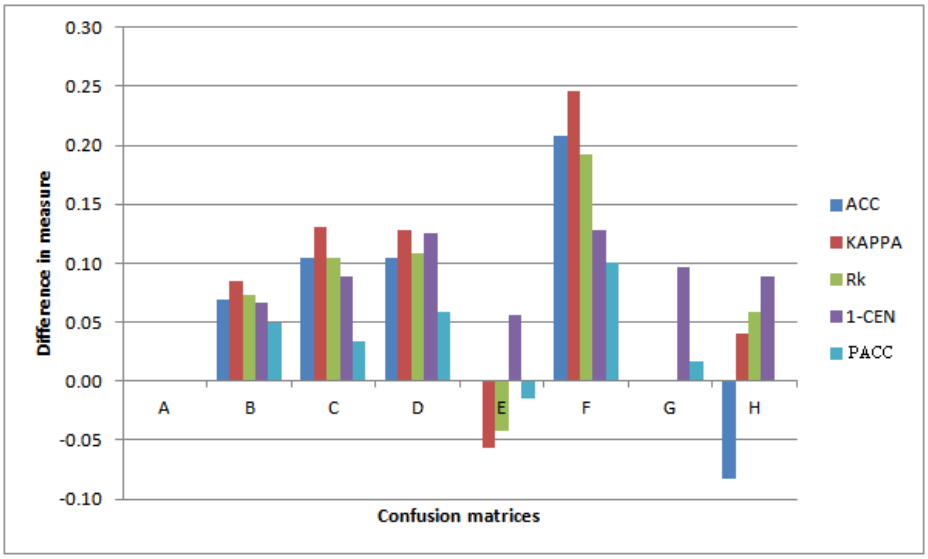

Fig. 2. Graph showing the difference in the values of performance measures in 4 and 6 .

considering different cases of confusion matrices. The results show that the proposed Pacc measure is relatively consistent with the accuracy measure and also is more discriminant than others. The Pacc measure was shown to be less affected by the scaling of data. Correlation coefficient and Macro-averaged F-measure can produce NaN and this does not necessarily happen when the performance is very low. Also, the interpretation of the results for Kappa statistic and correlation coefficient with values less than 0 is difficult. Likewise, we found that confusion entropy measure is not consistent.

Choice of a performance measure and its analysis can be domain/problem dependent. Also, analysis based on a single measure can be misleading as different measure can produce contrasting decision for the selection of a classifier. The measures may have specific biases and hence should be carefully used and analyzed. This follows that results of classification experiments should be accompanied by the confusion matrix.

As future work, we plan to investigate matrix normalization techniques to our proposed method which can be helpful for dataset with unbalanced class distribution. We also plan to formulate methods to compare confusion matrices of different sizes. Likewise, we would like to investigate methods to analyze performance where order of classification category is important.

Acknowledgments

This research was supported by National Institute of Health (GM-090453) grant.

References

1. Cumbaa, C.A., Jurisica, I.: Protein crystallization analysis on the world community grid. Journal of structural and functional genomics **11**(1) (2010) 61–69
2. Espndola, R.P., Ebecken, N.F.F.: On extending F-measure and G-mean metrics to multi-class problems. In: Data Mining, Text Mining and their Business Applications. (2005)
3. Özgür, A., Özgür, L., Güngör, T.: Text categorization with class-based and corpus-based keyword selection. In: 20th Inter. Conf. on Comp. and Inf. Sci. (2005) 606–615
4. Landgrebe, T., Duin, R.: Efficient multiclass roc approximation by decomposition via confusion matrix perturbation analysis. IEEE Trans on Patttern Analysis and Machine Intelligence **30** (2008) 810–822
5. Rees, G.S., Wright, W.A., Greenway, P.: Roc method for the evaluation of multi-class segmentation classification algorithms with infrared imagery (2002)
6. Yang, B.: The extension of the area under the receiver operating characteristic curve to multi-class problems. Volume 2. (2009) 463 –466
7. Wei, J., Yuan, X., Hu, Q., Wang, S.: A novel measure for evaluating classifiers. Expert Systems with Applications **37** (2010) 3799 – 3809
8. Gorodkin, J.: Comparing two k-category assignments by a k-category correlation coefficient. Computational Biology and Chemistry **28** (2004) 367 – 374
9. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. **45** (2009) 427–437
10. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H.: Assessing the accuracy of prediction algorithms for classification: An overview (2000)
11. Garca, S., Herrera, F.: An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. Journal of Machine Learning Research **9** (2009) 2677–2694
12. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7** (2006) 1–30
13. Perner, P.: How to interpret decision trees? In: Advances in Data Mining. Applications and Theoretical Aspects. Springer (2011) 40–55
14. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20** (1960) 37–46