Classifier performance

Dr. K.Vijayarekha
Associate Dean
School of Electrical and Electronics Engineering
SASTRA University, Thanjavur-613 401
Table of Contents

1. Classifier performance ............................................................ 3
   1.1 COMMON CLASSIFIER PERFORMANCE MEASURES ................................................................. 4
       1.1.1 Overall classification rate ................................................................................................. 4
       1.1.2 ROC Analysis .................................................................................................................. 5
   1.2 MISCLASSIFICATION COST AS PERFORMANCE MEASURE ............................................. 6
1. Classifier performance

Classifier performance evaluation is an important step in designing diagnostic systems. The purposes of performing classifier performance evaluation include:

1. To select the best classifiers from the several candidate classifiers
2. To verify that the classifier designed meets the design requirement, and
3. To identify the need for improvements in the classifier components.

In order to effectively evaluate classifier performance, a classifier performance measure needs to be defined that can be used to measure the goodness of the classifiers considered. In fault diagnostic system design, commonly used performance measures, such as accuracy and ROC analysis are not always appropriate for performance evaluation.

Figure (1) Typical Automated System Design Process for Diagnosis

Typical diagnostic system design involves several steps, including data preprocessing, feature extraction and selection, classifier design, and classifier performance evaluation, as shown in Figure 1. Classifier performance evaluation is an indispensable step in diagnostic system design. This is because the same classifier performs differently from application to application, i.e., classifier performance is problem specific. Given that no single classifier is always superior over others for all applications, common practice for designing classifier for a given problem, therefore, involves experimenting with many different classifiers, comparing their performance, and selecting the classifier (individual or combined) with the best performance. Obviously, in this design practice, classifier performance evaluation is essential. Performing classifier
Performance evaluation is required not only for selecting the best classifiers from the several candidate classifiers, but also for verifying that the classifier designed meets the design requirement and for identifying the need for improvements in the classifier components. Classifier performance generally refers to both computational performance and classification performance. In order to effectively evaluate classifier performance, a classifier performance measure has to be defined. A classifier performance measure is a single index that measures the Goodness of the classifiers considered. Depending on the design or application requirements, different problems may call for different performance measures to ensure that the classifiers considered can be properly compared and selected. Given a problem at hand, it is not always trivial to define a good performance measure so that the classifier performance can be accurately measured. In fact, defining/identifying a proper performance measure can be difficult for the cases when more comprehensive classifiers are required for the diagnostic system of a complex system.

Fault diagnostic systems have proved to be an effective technology in improving the reliability and in reducing the operating costs of various mechanical systems. There are great benefits for increasing the reliability/performance of fault diagnostics systems. However, as the systems become increasingly complex, designing a reliable fault diagnostic system becomes more difficult. Frequently, reasonable amount of design efforts result in only a small improvement in the performance of the diagnostic system designed. However, to the owner of the underlying system, even a small amount of improvement in reliability/performance of the diagnostic system can result in significant benefits. To uncover the subtle performance difference between one design and another, the performance measure used for classifier evaluation needs to be better defined to accurately represent the classifier performance. Diagnostic systems are typically designed to detect and isolate several different faults with different fault criticality. Consequently, classifier evaluation should take into account the difference between classifiers that have different misclassification costs for individual faults. Commonly used performance measures (accuracy and ROC analysis) are not always adequate for evaluating the performance of typical fault diagnostic systems. To overcome this problem, a more general performance measure, misclassification cost is utilized.

1.1 COMMON CLASSIFIER PERFORMANCE MEASURES
Common measures used for classifier performance evaluation include overall classification rate (accuracy) and ROC analysis. Descriptions of the two measures are given in this section. Advantages and limitations of the respective measures are discussed.

1.1.1 Overall classification rate
The most common measure of classifier performance is the overall classification rate (or alternatively its equivalent term, error rate, that is equal to one minus the overall classification rate). The overall classification rate, also called accuracy, is defined as the ratio of number of cases that are correctly classified over the total number of cases. Let CM be the M–by–M confusion matrix (where M is the number of classes), then the overall classification rate (OCR) is expressed as
\[ OCR = \frac{1}{N} \sum_{i=1}^{M} CM(i, i) \] (1)

Where \( N \) is the total number of test cases.

This single performance measure is fairly easy to compute. It is also suitable for all kinds of classifiers. The underlying assumption of the overall classification rate, however, is that the classification errors for all classes have equal cost consequences. Since this assumption rarely holds, the overall classification rate is often not an appropriate measure of classifier performance. Additional limitations of the overall classification rate as a performance measure include that it is sensitive to the unequal class size and that it is not reflecting the performance of the classifier across the entire range of possible decision thresholds. Some efforts have been made to make the accuracy measure useable for unequal error cost cases.

### 1.1.2 ROC Analysis

ROC Analysis

![ROC Curve](image)

**Figure 2**

ROC (short for Receiver Operating Characteristic or Relative Operating Characteristic) analysis is an established method of measuring diagnostic performance in various domains, especially in medical imaging studies. Originated from the field of signal detection to depict tradeoffs between hit rate and false alarm rate. ROC analysis and its associated indices have recently been extended for use in evaluating performance of two-class classifiers. The ROC space is a coordinate system that is used for visualizing classifier performance. In ROC space, the true positive rate, TPR, is plotted on the Y-axis and the false positive rate, FPR, is plotted on the X-axis, where the TPR is commonly referred to as “sensitivity” while (FPR) is called “specificity”. A point in ROC space corresponds to a (FPR, TPR) pair of a classifier. By varying the parameters of the Classifier, a series of points are obtained and a ROC curve is generated in ROC space by
connecting these points. Typical ROC curves are shown in Figure 2 where the three ROC curves represent three different classifiers. Typical ROC curves located in the upper-left corner in ROC space are better because they represent classifiers that have lower false positive rate and higher true positive rate than the classifiers below them.

ROC curves are a valuable technique for visualizing classifier behavior over a range of decision rules, therefore ROC curves are thus frequently used for selecting a suitable operating point, or a decision threshold, for the task at hand. However, when used for comparing or ranking classifiers based on their performance, evaluation based on ROC curves becomes more involved when the curves overlap. Use Figure 2 as an example where the three ROC curves correspond to three different classifiers A, B, and C. One can easily conclude that classifier C is better than or at least as good as the other two classifiers for all possible cost and class distributions since curve C dominates others in all range. Determining which of the two classifiers (A and B) is better, on the other hand, would not be so straightforward unless a specific performance requirement is given. For example, given the maximum number of false positive rate, one would simply draw a vertical line at the specified maximum FPR, and rank the classifiers based on TPR at the intersection of ROC curves with this vertical line. Similarly, one would use a horizontal line to rank the classifiers if the maximum true positive rate is given. When either the cost distribution or the class distribution is completely unknown and one wants to use a simple, single-quantitative index to represent the entire ROC curve, the area under the curve (AUC) is the most popular performance measure for ranking or comparing 2-class classifiers. The AUC provides a measure of performance that is not sensitive to the prior probability of class occurrence and measures the performance of the classifier across the entire range of decision thresholds. As noted above, all ROC analysis and its associated performance measures were developed for 2-class problems only, which is somewhat of a drawback of ROC analysis. This drawback restricts the ROC analysis from much wider applications.

1.2 MISCLASSIFICATION COST AS PERFORMANCE MEASURE

A general classifier performance evaluation poses the following requirements on performance measures:

First, performance measures need to be more accurate in order to uncover the subtle difference between one classifier and another. This is because small reliability improvement of diagnostic systems can bring significant benefits to the owner of the underlying system.

Second, performance measures need to be able to handle performance evaluation of multi-class classifiers. Most diagnostic systems are required to detect and isolate several different possible faults. With each of the classifier outputs representing one fault, most of diagnostic systems are multi-class classifiers.

Last, performance measures need to be able to represent criticality differences between faults, featuring in that the consequence cost of each fault is different.

These requirements make the commonly used performance measures discussed in the previous section inappropriate or inaccurate at best for evaluating fault diagnostic systems. More specifically, using the overall accuracy measure for evaluating classifiers with different
consequence cost of each fault would not be accurate due to the underlying assumption of equal
cost for misclassifying all faults. On the other hand, since ROC analysis and its associated
indices work for 2-class classifiers only, using the ROC analysis as the performance measure for
multi-class classifiers would be too complex and would lose visualization merits of ROC curves.
To address these problems, we use misclassification costs as a general classifier performance
measure for evaluating fault diagnostic systems. The misclassification cost (MC) is defined as
the product of each element of the normalized confusion matrix and the corresponding element of
the cost matrix and summing the results, expressed in Equation 2 as follows:

\[ MC = \sum_{i,j} cm(i,j) \cdot c(i,j) \]  \hspace{1cm} (2)

where, \( CM(i,j)' = \frac{CM(i,j)}{\sum CM(i)} \) is the normalized confusion matrix; the cost matrix and will be
discussed later in this section.

The misclassification cost defined in Equation 2 has been used by Margineantu and Dietterich
for designing cost-sensitive classifiers. One can see from Equation (2) that the overall
classification rate or accuracy is a special case of the misclassification cost. That is, when the cost
matrix has a value of 1 on its diagonal terms and zeros on all off-diagonal terms, the
misclassification cost becomes accuracy. In that sense, the misclassification cost measure is a
general form of the accuracy measure. The most significant advantages of the misclassification
cost measure is that it can be used for multi-class classifiers and can take care of classifiers with
different costs for different classes through proper definition of the cost matrix.

As shown in Equation 2, two elements of misclassification costs are the normalized confusion
matrix and the cost matrix. While the normalized confusion matrix can be readily calculated from
the confusion matrix of the classifier, precisely determining a cost matrix can be more difficult,
requiring domain expertise inputs or involvement of experts in different
areas from design to practice to field engineering.

A cost matrix is a matrix where each cell, \( C(i, j) \), represents the cost incurred for
misclassification, i.e., when a case is predicted to be in class \( j \) when in fact it belongs to class \( i \).
Based on this definition, it is obvious that all diagonal cells of a cost matrix should have a zero
value. We know that for a typical fault diagnosis problem, different faults have different
consequences. For example, in aircraft engine fault diagnosis, a fan fault will cause more severe
damage to the engine than a variable bleed valve (VBV) leak does. Hence, misclassifying a
critical fault (fan fault) as a non-critical fault (VBV fault) should have different cost consequence
from misclassifying a non-critical fault as a critical fault. Generally, cost of misclassifying fault “a”
as fault “b” should not be the same as that of misclassifying fault “a” as fault “c” if the
criticalities of faults “b” and “c” are different. Capturing this difference into performance
measures is the key for better evaluation of classifier performance. Thus, a full cost matrix is
generally a non-symmetric matrix. In modern mechanical system design, Failure Modes, Effects
and Criticality Analysis (FMECA) is used as a tool to gain a initial measure of system
reliability. The outcome of FMECA includes a list of failure modes of the system and associated
failure effects and failure criticality. The failure/fault criticality from FMECA provides valuable
information and can be used for estimating the cost matrix. Given the fault criticality ranking and
based on the notion that a more critical fault typically results in more costly effects, one could
derive a *heuristic* cost matrix such that the numbers of all cells in the same row are constant by dividing the *i*th ranking score by the sum of the ranking scores. However, such obtained cost matrix is not different from weighting each class based on its fault criticality, which is still not able to differentiate the misclassification costs between two different faults.

In order to estimate a full cost matrix, we propose to capture two basic notions: 1) the cost of misclassifying *i*th fault as *j*th fault is different from that of misclassifying *j*th fault as *i*th fault if *i* and *j* are different; and 2) the cost of misclassifying *i*th fault as *j*th fault is higher if the ordered criticality ranking of *j*th fault is further away from that of *i*th fault.