Receiver operating characteristic (ROC) analysis
Evaluating discrimination effects among decision support systems

Lena Kallin Westin
kallin@cs.umu.se

UMINF 01.18
ISSN-0348-0542

Department of Computing Science
Umeå University
SE-90187 Umeå
Sweden
Abstract

An overview of the usage of Receiver Operating Characteristic (ROC) analysis within medicine and computer science is given. A survey of the theory behind the analysis is given together with a presentation of how to create the ROC curve and different methods to interpret it. Both methods that rely on binormal distributions and methods that rely on distribution free methods have been mentioned.

A way to better know the quality of the measurements of sensitivity and specificity is presented. The quality measures and the QROC curve as is also included together with a discussion about optimal cut-offs and the connection to Bayesian decision theory. Results from earlier experiments and case studies will be used to exemplify the use of the ROC and QROC curves.
## Contents

1 The Theory of Receiver Operating Characteristic Curves 5
   1.1 Introduction ......................................................... 5
   1.2 Notations .......................................................... 5
   1.3 Sampling ............................................................ 7
      1.3.1 Estimates and standard errors for naturalistic sampling ..... 7
      Rule of thumb for $N_0$ ............................................. 8
      Example of a naturalistic sampling .............................. 8
   1.4 Different kind of tests ............................................. 9
      1.4.1 Test for legitimacy ............................................ 10
      Example of legitimacy of a test ................................. 10
      1.4.2 "Good" tests .................................................. 10
   1.5 The ROC curve ..................................................... 11
      1.5.1 Points on the ROC curve .................................... 12
      1.5.2 The optimal cut-off ........................................... 12
      Optimal cut-off points and Bayesian decision theory ........ 13
      1.5.3 The area under the ROC curve ............................. 15
      Comparing two areas ............................................. 15
      Comparing partial areas ......................................... 16
   1.6 The QROC curve .................................................... 16

2 Some Applications of Receiver Operating Characteristic Analysis 18
   2.1 Data sets used ..................................................... 18
   2.2 Methods used ..................................................... 19
      2.2.1 Using ROC/QROC curves the "naive" way .................. 19
      2.2.2 The preprocessing perceptron ............................... 19
      2.2.3 The fixed preprocessing perceptron ....................... 20
      2.2.4 Bayes decision theory ...................................... 20
   2.3 Results ............................................................. 20
      2.3.1 Comparing accuracies at optimal cut-offs ............... 20
      2.3.2 Comparing QROC-curves .................................... 21
      2.3.3 Comparing areas under two ROC-curves .................. 21

3 Summary and Conclusions ............................................. 23
Chapter 1

The Theory of Receiver Operating Characteristic Curves

1.1 Introduction

The Receiver Operating Characteristic (ROC) analysis comes from statistical decision theory [Green and Swets, 1966] and was originally used during World War II for the analysis of radar images. The first applications of this theory within the medical area occurred during the late 1960s. Today the ROC analysis is a widespread method in the medical field and many textbooks and articles, for example [Kraemer, 1992, Bamber, 1975, Metz, 1978, DeLong et al., 1988, Hanley, 1989, McClish, 1989, Armitage and Berry, 1994], have descriptions of it. From the computer science point of view, ROC analysis have been increasingly used as a tool to evaluate discriminate effects among different methods.

The ROC curve (described in Section 1.5) relies heavily on notations as sensitivity and specificity (see Section 1.2 for definitions) and these values depend on the specific data set. Even though the values for sensitivity and specificity in theory lie in the interval $[0, 1]$, in practice the borders are decided by the data set. If the QROC curve is considered instead (described in Section 1.6), a better way to compare different tests are given since all values have the same interval. This is accomplished by using two quality measures that transforms sensitivity and specificity values from different data sets to a comparable interval.

The paper is organised as follows; first, a presentation of the notations used and an introduction to the ROC and QROC curves as two methods to compare different medical tests will be given. This part is a summary of, among others, the work of [Metz, 1978, Hanley and McNeil, 1982, Hanley and McNeil, 1983, DeLong et al., 1988, McClish, 1989, Kraemer, 1992]. In the second part, the usage of ROC and QROC analysis to compare different decision support systems (or methods used in these systems) is described. Two data sets are used in the case studies, Down’s syndrome and diabetes data respectively. The results from different types of decision support systems will be analysed. Finally, summary and conclusions are presented.

1.2 Notations

ROC analysis is commonly used to evaluate medical tests and, in order to get an understanding of its use; this section describes some of the notations surrounding a medical test.

It is important to distinguish between disorder and diagnosis. A patient either has or has not a specific disorder during the period of testing. However, most medical tests are
compared to the diagnosis of the disorder and the measure is how well the result of the test corresponds to the diagnosis made. The diagnosis is not necessarily the correct answer but it is what the clinicians have to work with during the testing. Usually, diagnose is made by using one or more tests. If only one test is used this test is called the gold standard. A new test that is found to work better than the current gold standard might replace it and become the new gold standard in the future.

In the book *Evaluating medical tests* [Kraemer, 1992], the author claims that the following assumptions are made and must be fulfilled in order to be able to do an evaluation of a medical test:

1. **The period of testing must be defined so that throughout this period the patient either has the disorder or not.**

2. **The purpose of a medical test is to help determine whether or not a patient has a specific disorder during the period of testing.**

3. **The diagnosis of the disorder used to evaluate a medical test must be clinically valid.**

The first step when performing a medical test is to create a sample group of patients. Different ways to create sample groups are described in Section 1.3. In the following, it is assumed that a sample group has been chosen, diagnosed and tested.

Let $p_i$ be the probability that patient $i$ will get a positive diagnosis (i.e., the patient is ill) and $q_i$ be patient $i$’s probability of a positive test. The prevalence, $P$, of the positive diagnosis in the population is theoretically $P = \text{mean}(p_i)$. The level of the test, $Q$, is $Q = \text{mean}(q_i)$. We also define $P' = 1 - P$ and $Q' = 1 - Q$.

<table>
<thead>
<tr>
<th>Test result</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Negative</td>
<td>FP</td>
<td>TN</td>
</tr>
<tr>
<td></td>
<td>$Q$</td>
<td>$Q'$</td>
</tr>
</tbody>
</table>

Table 1.1: Relations between the measurement probabilities of the outcome, prevalence, and level of a test defined in the text.

In general, four possible decisions and two types of errors are made when comparing a test result with a diagnosis as shown in Table 1.1. If both diagnosis and test are positive, it is called a true positive. The probability of a TP to occur is estimated by counting the true positives in the sample and divide by the sample size. If the diagnosis is positive and the test is negative it is called a false negative (FN). False positive (FP) and true negative (TN) are defined similarly.

The values described are used to calculate different measurements of the quality of the test. The first one is sensitivity, SE, which is the probability of having a positive test

---

1Kraemer defines clinically valid as the case when patients that are more likely to have a positive diagnosis also are more likely to have the disorder.

2Note that during a test the prevalence for the considered disorder is always unknown for all that is quantified concerns the diagnosis not the disorder.
among the patients who have a positive diagnosis.

\[ \text{SE} = \frac{TP}{TP + FN} = \frac{TP}{P}. \]

Specificity, SP, is the probability of having a negative test among the patients who have a negative diagnosis.

\[ \text{SP} = \frac{TN}{FP + TN} = \frac{TN}{P'}. \]

Efficiency is defined as \( \text{EFF} = \frac{TP}{TP + TN} \). All three measurements will be used frequently in this report. Two other measurements that can be used are the predictive value of a positive test, \( \text{PVP} = \frac{TP}{TP + FP} = \frac{TP}{Q} \) and the predictive value of a negative test, \( \text{PVN} = \frac{TN}{TN + FN} = \frac{TN}{Q'} \).

1.3 Sampling

To be able to perform a medical test, a sample of the population must be collected to perform the test on. There are essentially three different ways to create a sample [Kraemer, 1992]:

**Naturalistic sampling**

The sampling is done by gathering a number \( N_0 \) of people that are representative of the population in interest. Each patient in the sample are both diagnosed and tested. This is a natural way to do a sampling but difficult to perform in practice.

**Retrospective sampling**

If the test is costly and/or risky, one might not want to test every person in the sample. In retrospective sampling, \( N_0 \) people are drawn from the population and all are diagnosed. This sample is called the screening sample. From this sample, a random sample of \( N_1 \) people with positive diagnosis and a random sample of \( N_2 \) people with negative diagnosis is drawn. Then only these \( N_1 + N_2 \) are tested.

**Prospective sampling**

Prospective sampling is done when a number of patients \( N_0 \) is drawn from the population and tested. Then a random sample of \( N_1 \) patients with a positive test is selected together with a random sample of \( N_2 \) patients with a negative test. Only these \( N_1 + N_2 \) patients are diagnosed.

From a computer science point of view, a data file is received containing numerical values representing symptoms for a specific disease together with the diagnose. Tests are performed on the data, i.e. apply an algorithm on the data to receive an output that symbolises the diagnosis. Mostly, all cases are tested and therefore naturalistic sampling is used (if the clinicians have created the data set in a representative way). It can also be viewed as retrospective sampling is used. However, in that case the value of \( N_0 \) is unknown. In the rest of this paper, it is assumed that naturalistic sampling is used. For more information about retrospective and prospective sampling, see [Kraemer, 1992].

1.3.1 Estimates and standard errors for naturalistic sampling

The estimations of the prevalence, \( \hat{P} \), the level, \( \hat{Q} \) together with the estimates for TP, FN, FP, TN, and EFF are all unbiased and they all follow the same pattern [Kraemer, 1992]:

\[ \text{SERR}(\hat{X}) = \sqrt{\hat{X}(1 - \hat{X})}/N_0, \]
where $X$ is one of the measures mentioned earlier. The standard error can be estimated by replacing $X$ with $\hat{X}$ in the formula.

Sensitivity and specificity becomes a bit harder. The estimates are

$$\hat{SE} = \frac{TP}{\hat{P}} \quad \text{and} \quad \hat{SP} = \frac{TN}{\hat{P}^0}$$

But these estimates are not unbiased\(^3\). The bias depends on the size of the sample $N_0$ as well as on $P$. This bias gets smaller as $N_0$ gets larger and if it is large enough compared to $P$ the problem can be avoided. In the case of a low-risk population, $N_0$ must be larger than in a higher-risk population. If $N_0$ is large enough, the standard errors can be estimated in the following way:

$$SERR(\hat{SE}) = \sqrt{\frac{SE(1 - SE)}{(N_0 \cdot P)}}$$
$$SERR(\hat{SP}) = \sqrt{\frac{SP(1 - SP)}{(N_0(1 - P))}} \quad (1.1)$$

**Rule of thumb for $N_0$**

If a table is made, as done in Table 1.2 for our example, each number in the marginal positions must be at least 10 [Kraemer, 1992]. If both the prevalence and level of the test is 0.5, $N_0$ must be $\geq 20$. If the prevalence is low, e.g. 0.01, $N_0$ must be at least 1000. With this rule of thumb the estimators will be reasonably unbiased and the estimated standard errors reasonably accurate.

**Example of a naturalistic sampling**

If naturalistic sampling is used, the resulting population can be described as in Figure 1.1. Each patient has either a positive or a negative diagnosis shown as plus and minus signs in the figure. The test can be either positive or negative and is illustrated by the colour of the signs (black for positive and white for negative). If the four different types of signs is counted and the numbers filled in, a table is created as shown in Table 1.2.

Figure 1.1: A sample population of $N = 95$ patients. The minus signs denote patients with a negative diagnosis and the plus signs denotes patients with a positive diagnosis. The colour shows the result of a test. White is a negative test and black is a positive test.

---

\(^3\)With a repeated sampling from a (low-risk) population there is some probability to get a sample where none of the patients has a positive diagnosis ($\hat{P} = 0$). Since the SE cannot be estimated from such a sample some rule to handle that case must be defined. Whatever rule, there is a bias introduced. (Often, there is an underestimation of SE)
### Test result

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>30</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>33</td>
</tr>
<tr>
<td>Negative</td>
<td>20</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95</td>
</tr>
</tbody>
</table>

Table 1.2: Table with numbers from the example in Figure 1.1 filled in.

The estimates of the values and their standard errors are as the following:

\[
\begin{align*}
TP &= \frac{30}{95} = 0.316(\pm 0.048) & P &= \frac{33}{95} = 0.347(\pm 0.049) \\
FN &= \frac{3}{95} = 0.032(\pm 0.018) & Q &= \frac{50}{95} = 0.526(\pm 0.051) \\
FP &= \frac{20}{95} = 0.211(\pm 0.042) & SE &= \frac{30}{33} = 0.909(\pm 0.050) \\
TN &= \frac{42}{95} = 0.442(\pm 0.051) & SP &= \frac{42}{62} = 0.677(\pm 0.059)
\end{align*}
\]

### 1.4 Different kind of tests

If there is a strong correspondence between a test and a clinically valid diagnosis, it can be concluded that the test is a valid indicator of the presence of the disorder. The correlation between the probability of a positive diagnose and a positive test for a patient is described by \( \rho \). The following is known [Kraemer, 1992]:

\[
TP = \text{mean}(p_iq_i) \text{ or } \quad TP = PQ + \rho \sigma_p \sigma_q
\]

and in the same way:

\[
\begin{align*}
TN &= P'Q' + \rho \sigma_p \sigma_q \\
FP &= P'Q - \rho \sigma_p \sigma_q \\
FN &= PQ' - \rho \sigma_p \sigma_q
\end{align*}
\] \( (1.2) \)

A test is *ideal* when there is a perfect match between test and diagnosis. This is only possible if the prevalence \( (P) \) is the same as the level \( (Q) \) of the test, i.e.,

\[
\begin{align*}
TP &= P \quad FN = 0 \\
FP &= 0 \quad TN = P'
\end{align*}
\] \( (1.3) \)

In a *random test* there is no correspondence at all (i.e., \( \rho = 0 \)) and the following equations hold:

\[
\begin{align*}
TP &= PQ \quad FN = PQ' \\
FP &= P'Q \quad TN = P'Q'
\end{align*}
\] \( (1.4) \)

In a *legitimate test* there is at least some correspondence \( (\rho > 0) \)

\[
\begin{align*}
TP &> PQ \quad FN < PQ' \\
FP &< P'Q \quad TN > P'Q'
\end{align*}
\] \( (1.5) \)

If a test is legitimate or random, a patient with a positive test are at least as likely to have a positive diagnosis as those with a negative test. If a test has a negative correspondence it can be made legitimate by swapping the test criteria.
1.4.1 Test for legitimacy
To check if a test is good or not, the first criteria is that it should at least be legitimate, i.e., fulfil the equations in Eq. 1.5. The bigger difference in the equations, the better test (i.e., the test is closer to the equations in Eq. 1.3). The Chi-square test is used to check this.

To use the common Chi-square test one has to compute a test statistic and compare it to a value. If the statistic is above a certain value (3.85 at 5% significance level, 6.63 at 1% significance level, and 10.83 at 0.1% significance level) the medical test is a legitimate one. The test statistic can be computed as follows [Kraemer, 1992]:

\[ \chi^2 = N_0 \cdot \kappa(1,0) \cdot \kappa(0,0), \]

where

\[ \kappa(1,0) = (SE - Q)/Q' \]
\[ \kappa(0,0) = (SP - Q)/Q \]

(1.6)

\( \kappa(1,0) \) and \( \kappa(0,0) \) are called quality indices for the test, see also Section 1.4.2.

Example of legitimacy of a test
Now, it can be checked if the earlier example (see Section 1.3.1) is legitimate. We get the following values:

\[
\begin{align*}
TP &= 0.316 & PQ &= 0.183 & FN &= 0.032 & PQ' &= 0.164 \\
FP &= 0.211 & P'Q &= 0.343 & TN &= 0.442 & P'Q' &= 0.310
\end{align*}
\]

Equation 1.5 is fullled but it must be shown statistically as well.

\[
\begin{align*}
\kappa(1,0) &= (0.909 - 0.526)/0.474 = 0.808 \\
\kappa(0,0) &= (0.677 - 0.474)/0.526 = 0.387 \\
\chi^2 &= 29.72 (p < 0.001)
\end{align*}
\]

The conclusion can be drawn that the test that produced the resulting population shown in Figure 1.1 was a legitimate one. However, the conclusion that the test is a good one cannot be drawn. For a test to be good it must at least be legitimate but there are many tests that fulfil that. In the next section, it will be discussed how to recognise a good test.

1.4.2 "Good" tests
When a legitimate test is created, one can start investigating if the test is "good". By using Equation 1.2, sensitivity and specificity are defined as:

\[
\begin{align*}
SE &= TP/P = Q + \rho \sigma_p \sigma_q/P \\
SP &= TN/P' = Q' + \rho \sigma_p \sigma_q/P'
\end{align*}
\]

For a random test (\( \rho = 0 \)) the sensitivity equals the level of the test and the specificity equals the complement of the level of the test. For a legitimate test the values are higher. The ideal value for both SE and SP is 1.0, which may not be achievable. If a report of sensitivity and specificity of a test does not include the level of the test, there is no indication of the quality of the test. Kraemer [Kraemer, 1992] introduces two quality measures to use, redefined in Equation 1.6.

If the sensitivity is 99.9% and the level of the test is 99.9% the sensitivity is of zero quality (\( \kappa(1,0) = 0 \)). On the other hand a specificity of 70% for a test with level 99% may
be an outstanding test ($\kappa(0, 0) = 0.697$). The choice of the "best" test should depend on the values of $\kappa(1, 0)$ and $\kappa(0, 0)$ - not on the values of sensitivity and specificity but on their quality.

Our example from Figure 1.1 and Table 1.2 had the level $Q = 0.526$, the sensitivity $SE = 0.909$ and specificity $SP = 0.677$. This gives us $\kappa(1, 0) = 0.81$ and $\kappa(0, 0) = 0.39$. The quality indexes confirm the suspicion that the sensitivity has better quality than specificity in this case. $\kappa(0.5, 0)$ is a measure that gives the overall quality of a test if efficiency ($EFF = TP + TN$) is considered. It is zero for a random test and one for an ideal test. It is defined as [Kraemer, 1992]:

$$\kappa(0.5, 0) = \frac{PQ\kappa(1, 0) + P'Q\kappa(0, 0)}{PQ + P'Q}$$

In our example $\kappa(0.5, 0) = 0.526$.

### 1.5 The ROC curve

A decision support system gives output in the interval $[0, 1]$ where 0 denotes a negative and 1 denotes a positive diagnosis. By introducing a cut-off somewhere in the interval the output is binarised and compared to the true diagnosis given by a clinician. Each cut-off corresponds to a point on a ROC curve. The ROC curve has the sensitivity plotted vertically and the reversed scale of the specificity on the horizontal axis. The scale of the horizontal axis is also called the false positive rate$^4$. The sensitivity and specificity, and therefore the performance of the system, vary with the cut-off. If several tests is performed on the same sample, ROC curves can be used to compare their performance. Another way to use ROC curves is to see the performance of a decision support system. The correspondence between cut-off and performance can be shown by an example.

![Figure 1.2: Three pairs of distributions with large, medium, and small overlap respectively.](image)

Assume that the result of the decision support system forms two normal distributions, one for the healthy patients and one for the ill patients. The threshold is placed at different positions to divide the distributions. The sensitivity and specificity is calculated for each position and the resulting points are plotted as a ROC curve. The curve gives a picture of the performance of the system. In Figure 1.2, three examples with ten different decision thresholds are shown. The overlap between the distributions is largest in the first example and smallest in the third.

When plotting the sensitivity versus 1 - specificity values for these cut-offs we get the curves shown in Figure 1.3. The point marked by an arrow corresponds to the sensitivity and specificity at the cut-off value 0.222 in the distributions with the largest overlap.

$^4$It is called the false positive rate since $1 - SP = 1 - \frac{TN}{P'} = \frac{FP + TN - TN}{FP + TN} = \frac{FP}{FP + TN}$
Figure 1.3: The corresponding ROC-curves for three pairs of distributions with large, medium, and small overlap respectively.

### 1.5.1 Points on the ROC curve

Each point on the curve corresponds to a specific pair of sensitivity and specificity and the complete curve\(^5\) gives an overview of the overall performance of a test. When comparing ROC-curves of different tests, good curves lie closer to the top left corner and the worst case is a diagonal line (shown as a dashed line in Figure 1.3). There are methods for estimating confidence intervals for ROC curves as well. For more information see e.g. [Reiser and Faraggi, 1997, Schäfer, 1994, Zou et al., 1997].

When concentrating on a specific point on a ROC curve it might be interesting to find the confidence region for the sensitivity and specificity. The standard error is calculated by Equation 1.1. However, as said earlier, it is important to have large enough sample. The confidence region, CR, is then given by \( CR = X \pm t_{a(2),n-1} \times SERR(X) \) where \( X \) is the estimated sensitivity (or specificity). If the number of patients, \( P \) or \( P' \) is small it is better to use a binomial distribution instead [Beck and Shultz, 1986].

Sometimes, there are two curves present and then it is interesting to compare sensitivity (or specificity) at a given point on the ROC curves. If the data used to construct the two curves are the same we have paired samples and only the samples that disagree with each other are used. The test statistic is given by [Beck and Shultz, 1986]:

\[
\chi^2 = \frac{(|t - f| - 1)^2}{t + f},
\]

where \( t \) is the number of samples correctly classified by the first decision support system but incorrectly by the second. The samples correctly classified by the second system but not by the first is denoted \( f \). The statistic is corrected for small sample sizes and should be compared to the \( \chi^2 \)-table with one degree of freedom.

If the samples are not paired, either the \( \chi^2 \)-test or the Fischer exact solution (when high sensitivity and specificity occurs with small sample sizes) are used. For more details about these tests, see [Beck and Shultz, 1986].

### 1.5.2 The optimal cut-off

Often the ROC analysis is used to find an optimal cut-off value for use in decision-making. Usually a clinician wants a specificity of 95% and tries to find the cut-off for that. However,\(^6\)

\(^5\)Or more correctly the area beneath it, see Section 1.5.3
in some cases it is almost as important to have a high sensitivity, e.g. in the case of Down’s syndrome. In Figure 1.4, sensitivity and specificity plots are shown for the example with artificial data described earlier in this section. The third plot shows the total performance for the data and it can be seen that the best performance is accomplished by choosing the cut-off where the specificity and sensitivity crosses. However, as mentioned before, the optimal cut-off depends on the situation.

![Sensitivity, Specificity, Total Performance Plots](image)

Figure 1.4: In the first row, graphs of the sensitivity and the specificity is shown. The graph in the second row shows the total performance for the examples mentioned in the text. Note that, in this example, the maximum of the total performance is placed were the sensitivity and the specificity crosses.

Remember also the discussion in Section 1.4.2 about the quality of the values. In Table 1.3, an example of different sensitivities, specificities and their corresponding quality values are shown. The corresponding ROC curve is shown in Figure 1.5. If the test with the best sensitivity is needed, the cut-off value 0.1 should be chosen which gives a sensitivity of 100% and a quality measure of 100%. If the best specificity is wanted, the figure shows that the highest value is 96.0% at the cut-off 0.9 and quality 83.6%. If the overall best test is sought for, the test with cut-off 0.5 should be chosen. In this case sensitivity is 78.7%, specificity 86% and the quality measure is 64.8%. The conclusion is the “the optimal cut-off” depends on what will be done with the test later on.

**Optimal cut-off points and Bayesian decision theory**

Bayesian decision theory works with predictive values and ROC/QROC analysis with values of sensitivity and specificity. Let $D^+$ and $D^-$ represent a positive and a negative diagnose respectively. In the same way let $T^+$ and $T^-$ denote the test result. The posterior probability that the diagnose is positive given that the test is positive, $P(D^+ | T^+)$ is the

---

6The example is from a data set for Down’s syndrome which will be described later on.
Table 1.3: Table with sensitivities, specificities and quality measures for a data set. All values except cut-off are given in %.

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>PVN</th>
<th>PVP</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>(\kappa(1, 0))</th>
<th>(\kappa(0, 0))</th>
<th>(\kappa(0.5, 0))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Insufficient sample size (test not legitimate)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>51.5</td>
<td>100.0</td>
<td>100.0</td>
<td>11.3</td>
<td>100.0*</td>
<td>5.8</td>
<td>11.0</td>
</tr>
<tr>
<td>0.2</td>
<td>60.5</td>
<td>95.2</td>
<td>97.9</td>
<td>40.0</td>
<td>90.2</td>
<td>23.4</td>
<td>37.2</td>
</tr>
<tr>
<td>0.3</td>
<td>69.0</td>
<td>88.5</td>
<td>91.5</td>
<td>61.3</td>
<td>76.2</td>
<td>39.8</td>
<td>52.3</td>
</tr>
<tr>
<td>0.4</td>
<td>78.8</td>
<td>84.3</td>
<td>83.7</td>
<td>87.7</td>
<td>67.6</td>
<td>58.9</td>
<td>62.9</td>
</tr>
<tr>
<td>0.5</td>
<td>84.1</td>
<td>81.1</td>
<td>78.7</td>
<td>86.0</td>
<td>61.1</td>
<td>69.1</td>
<td>64.8*</td>
</tr>
<tr>
<td>0.6</td>
<td>84.6</td>
<td>75.9</td>
<td>70.2</td>
<td>88.0</td>
<td>50.2</td>
<td>70.2</td>
<td>58.5</td>
</tr>
<tr>
<td>0.7</td>
<td>86.4</td>
<td>72.3</td>
<td>63.1</td>
<td>90.7</td>
<td>42.9</td>
<td>73.6</td>
<td>54.2</td>
</tr>
<tr>
<td>0.8</td>
<td>89.2</td>
<td>70.7</td>
<td>58.9</td>
<td>93.3</td>
<td>39.5</td>
<td>79.1</td>
<td>52.7</td>
</tr>
<tr>
<td>0.9</td>
<td>91.5</td>
<td>65.5</td>
<td>46.1</td>
<td>96.0</td>
<td>28.7</td>
<td>83.6*</td>
<td>42.7</td>
</tr>
<tr>
<td>1</td>
<td>Insufficient sample size (test not legitimate)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the same way, the probability of a negative diagnose when the test is negative \(P(D-|T-)\) is the negative predictive value (PVN):

\[
P(D-|T-) = \frac{P(T-|D-)P(D-)}{P(T-)} = (SP \cdot P')/Q' = TN/Q' = PVN.
\]

The predictive values depend on the prevalence and if we scale the values as done with sensitivity and specificity we notice that the quality values for the specificity is the same as the positive predictive value and the quality values for sensitivity is the same as for the negative predictive value.

\[
\begin{align*}
\kappa(1, 0) &= (SE - Q)/Q' = (PVN - P')/P \\
\kappa(0, 0) &= (SP - Q')/Q = (PVP - P)/P'
\end{align*}
\]  

This means that if we maximise \(\kappa(1, 0)\) we have at the same time found the maximum sensitivity and the maximum negative predictive value.

![ROC-curve](image)

Figure 1.5: The ROC-curve for the values shown in Table 1.3.

It has been shown that the Bayesian decision procedure has optimal performance when minimising the risk for discriminating wrong between classes [Duda et al., 2001]. If we
consider the minimum-error-rate classification in the two-category case we have a bayesian discriminant function that is the difference between the two posterior probabilities. This is expressed as (in the notation of [Duda et al., 2001]):

\[ g(x) = g_1(x) - g_2(x) = P(\omega_1|x) - P(\omega_2|x). \]

If \( g(x) > 0 \) we should decide that \( x \) belongs to the \( \omega_1 \), otherwise \( \omega_2 \). This means that the border between the classes (the optimal cutoff) is found where \( g(x) = 0 \). To be able to use Bayes method, one need to know the prior probabilities \( P(\omega_i) \) of the disease as well as the conditional densities \( p(x|\omega_i) \). If these are not known they can be estimated from the data set. The performance of the Bayesian decision procedure therefore depends on the quality of these estimates and can not be optimal if the wrong distribution is assumed or the estimates are poor.

The optimal cut-off used in this article is based on the efficiency of a test, \( EFF = TP + TN \), and is built on Cohen’s kappa and is found by maximising

\[ \kappa(0.5, 0) = \frac{PQ\kappa(1, 0) + P'Q\kappa(0, 0)}{PQ' + P'Q}. \]

1.5.3 The area under the ROC curve

The total area under the ROC-curve is a measure of the performance of the diagnostic test since it reflects the test performance at all possible cut-off levels. The area lies in the interval \([0.5, 1]\) and the larger area, the better performance. Assume that a high value from the method indicates that diagnosis is positive and a low value indicates that diagnosis is negative. The area is then a measurement of the probability that the distribution of the positive diagnosis is statistically larger than the distribution of the negative diagnosis. Many articles discuss the area and how to calculate and interpret it, e.g. [Bamber, 1975, Beck and Shultz, 1986, DeLong et al., 1988, Hanley and McNeil, 1982, Hanley and McNeil, 1983]. In this paper, much of the mathematical/statistical derivations are omitted; details can be found in the references.

In experiments, there is usually only a finite set of points on the ROC-curve. Therefore it is only possible to find an approximation of the area under the curve. In Figure 1.6 an example of this is shown. It is obvious that the more points there are, the better estimate of the curve and area we get. There are several ways to calculate the area under a ROC curve. First, the trapezoidal rule can be used but gives an underestimation of the area. Second, it is possible to get a better approximation of the curve by fitting the data to a binormal model with maximum-likelihood estimates. After that it is possible to get a better estimate of the area. This is done, for example, in the program Rockit [Rockit, 2002]. A third way to calculate the area is to use the Mann-Whitney U statistic (also known as the non-parametric Wilcoxon statistic). That is, no assumptions on the distributions of the data are done since Wilcoxon is a distribution-free statistic [Bamber, 1975, Hanley and McNeil, 1982]. The program Rockit also presents a Wilcoxon area-estimation.

Comparing two areas

Often, there is a need to compare different methods applied on the same data set and compare the ROC curves in order to determine which method is the best. In such cases it is important to take the correlation between the areas that is induced by the data into account. If this is done, the standard error is reduced and the power of the comparison increases. In other words, it is easier to detect differences in areas if the correlation is counted for.
When comparing two areas the critical ratio is defined by \[ z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}}. \]

\( A_1 \) and \( A_2 \) are the two areas and \( SE_1 \) and \( SE_2 \) the corresponding standard errors and \( r \) is the quantity representing the correlation between the two areas due to working on the same set of data. If the program Rockit \[ \text{[Rockit, 2002]} \] is used estimates of the area and correlation factor is given. In \[ \text{[DeLong et al., 1988]} \] there is a description of how to make a nonparametric comparison of areas under two correlated ROC curves.

It is important to point out that a non-significant difference between areas for two methods does not imply equivalence between the methods. In order to say that two methods are equivalent, a definition of equivalence must be made. The meaning of a non-significant difference is that there is insufficient evidence to show a difference smaller than a specific amount.

Comparing partial areas

Some papers \[ \text{[McClish, 1989, Obuchowski and McClish, 1997, Zou et al., 1997]} \] discuss the fact that when someone is comparing the full areas, equal weight is given to all false positive error rates. Often the clinician is interested only in differences with in a specified interval, e.g., the area where the specificity lies in the interval \([0.8, 1]\). In these cases comparisons of partial areas should be done instead. In \[ \text{[McClish, 1989]} \], it is shown how to analyse partial areas under the assumption that the data is binormal. In the article methods are available for both dependent and independent data.

1.6 The QROC curve

A QROC curve has the quality measurement \( \kappa(0,0) \) on the horizontal axis and \( \kappa(1,0) \) on the vertical. In Figure 1.7 the ROC curve for our example (see also Table 1.3) in the previous section is shown together with its corresponding QROC curve.

In a QROC curve, one can immediately see which test that have the optimal sensitivity. It is the marker at the highest position (labelled 0.1). The marker most to the right is the one with the optimal specificity (labelled 0.9). Finally, the test with the optimal overall efficiency is the marker closest to the upper right corner (labelled 0.5).

All legitimate tests have quality values that are positive. The point \((0,0)\) corresponds to a random test and the point \((1,1)\) corresponds to an ideal test. Any point inside the
Figure 1.7: An example of a ROC curve (left) and its corresponding QROC curve (right). The numbers in the QROC are cut-off values used to produce that dot.

QROC (i.e., not on one of the axis) will be statistically significant, if the sample size is large enough, for details see [Kraemer, 1992].
Chapter 2

Some Applications of Receiver Operating Characteristic Analysis

In this chapter, it will be shown how ROC (and QROC) analysis can be, and are typically, used when comparing different methods for decision support systems. In the first section, the two data sets used are described. In the second section, four different methods to build decision support systems are described together with an explanation of how they are used in our case study. The results are described in the last section.

2.1 Data sets used

The first data set is data for the Down’s syndrome. The inputs to the system were the mother’s age (in years) and measurements of the hormones AFP and free $\beta$-hCG. Our data are collected in Europe and USA and have 300 non-Down and 282 Down cases used in the learning phase, and 150 non-Down and 141 Down cases in the testing phase. See [Kallin et al., 1998] for further details of the experiment and data used.

The second case study is built on data for Pima-Indians-Diabetes. The data were originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases [Blake and Merz, 1998]. Eight inputs are used; number of times pregnant, plasma glucose concentration a 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-Hour serum insulin ($\mu$U/ml), body mass index (kg/m²), diabetes pedigree function, and age (years). Our data consists of 375 non-Diabetes and 201 Diabetes cases used in the learning phase, and, respectively, 125 non-Diabetes and 67 Diabetes cases in the testing phase. A data set where all missing data are set to 0.5 will be used, see [Eklund and Kallin Westin, 2002] for details about the data set and its missing data.

The first data set is special in the sense that the test is done during the pregnancy and are ground for an early amniocentesis if the “diagnose” is “Down”. Since an early amniocentesis can be dangerous for the foetus it is important to keep the number of false positives low. This restriction is not present in the second data set.

To minimise effects from dividing the data sets in an unequal way, ten randomly divided sets are created for each of the case studies and the test is performed on each set. The results are then added together and the average result is presented. Each of the ten sets contains exactly the same number of individual cases. The difference lies in the dividing of the large set into training and test sets. The size of the test and training sets are also the same in all ten cases.
2.2 Methods used

2.2.1 Using ROC/QROC curves the "naive" way

The first type of decision support system is perhaps a "naive" way to use ROC/QROC curves. First, all data is normalised to the [0, 1]-interval with the formula

\[ x^* = \frac{(x - \text{min})}{(\text{max} - \text{min})}. \]

Max and min values are taken from the training set and if some values in the test set lies outside the interval, they are truncated to the interval limits. Second, QROC analysis is made on each of the inputs (with all data from the training set) to find the overall optimal cut-off for the input (i.e. \( \kappa(0.5, 0) \)). The cut-offs obtained are used on the inputs in the test set. The input value is set to zero if it is below the cut-off and to one otherwise. Last, a diagnosis is made according to the rule; if two or more of the inputs are set to one set the output (diagnosis) to one, otherwise set it to zero. This will give us a decision support system that looks at all inputs one by one. Each input will give a response; i.e. zero for healthy and one for ill. Finally the results are added together into a decision that will be compared to the diagnosis in the data file.

The average value of the accuracy and the cut-offs chosen for the ten different test sets were recorded. When considering the optimal cut-off for each of the ten different data sets it was noticed that similar values were obtained. In Table 2.1 the minimum, maximum and mean value for the different inputs are shown. The second parameter for Down’s syndrome had all values but one above 0.86.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Down’s syndrome</th>
<th>Diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>1</td>
<td>0.5455</td>
<td>0.6364</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.1286</td>
<td>0.1484</td>
</tr>
</tbody>
</table>

Table 2.1: Mean, minimum and maximum values of the optimal cut-offs for the 10 different data sets. Down’s syndrome data is shown to the left and diabetes data to the right.

2.2.2 The preprocessing perceptron

The preprocessing perceptron can be seen as a single-layer perceptron with a preprocessing layer added. This preprocessing layer usually consists of sigmoidal functions. These sigmoidal functions are trained together with the net using the gradient descent method in a way similar to the backpropagation algorithm. The output from a preprocessing perceptron is given by

\[ o = g[0, 1] \left( \sum_i w_i g[\alpha_i, \beta_i](x_i) \right), \]
where the sigmoid function is given by,

\[ g(\alpha, \beta) = \frac{1}{1 + e^{-\beta(x-\alpha)}} \]

\(x_i\) is the inputs and \(w_i\) the weights in the network. For more information about the preprocessing perceptron see, e.g., [Kallin, 1998, Kallin Westin, 2002].

As in the previous case, each setting is trained 10 times, on different data sets, and the mean value of the accuracy and the resulting \(\alpha:\)'s of these are reported.

### 2.2.3 The fixed preprocessing perceptron

The only difference compared to the preprocessing perceptron is that the preprocessing layer is fixed to values that are optimal for each individual input. The cut-off has been chosen with the help of a QROC curve. The obtained values are fixed during the training, i.e., only the weights are trained.

### 2.2.4 Bayes decision theory

A minimum-error-rate bayesian classifier has been implemented according to the formulas in [Duda et al., 2001]. The prior probabilities and the conditional densities has been estimated from the training data set and used in the test data set.

### 2.3 Results

The result from the first method is a binary decision value, which is compared to the diagnosis to get an accuracy number. The second method has a value in the interval \([0, 1]\) as output and the choice of cut-off between the positive and negative diagnoses decides the performance of the system. QROC was used to find the best cut-off for the second method.

#### 2.3.1 Comparing accuracies at optimal cut-offs

The first way to compare these methods is to compare the accuracies. The results are shown in Table 2.2.

<table>
<thead>
<tr>
<th></th>
<th>Naive</th>
<th>PP</th>
<th>Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Down</td>
<td>0.5515</td>
<td>0.8186</td>
<td>0.7852</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.662</td>
<td>0.7755</td>
<td>0.7469</td>
</tr>
</tbody>
</table>

Table 2.2: Mean values for accuracies for the different methods, 10 data sets has been used.

In both cases the preprocessing perceptron outperforms the naive decision support system (\(p < 0.001\) for Down and \(p = 0.012\) for Diabetes data). It is, however, interesting to see that in neither case there is any significant difference between the preprocessing perceptron and the bayesian classifier. In the naive case the Diabetes data has significantly better performance than the Down data (\(p = 0.013\)) while it is not true in the preprocessing perceptron case (\(p = 0.267\)).
2.3.2 Comparing QROC-curves

To illustrate how QROC curves can be used to analyse results, the results of two data sets from the Down data (using the preprocessing perceptron) will be used. In Figure 2.1 the ROC and QROC curves for data set five and ten is shown. The optimal cut-off is shown together with information about sensitivity, specificity and quality values.

Figure 2.1: ROC and QROC curves corresponding to data set 5 and 10 from Down data (with preprocessing perceptron used).

It can be seen in the QROC curve that the optimal cut-off is a point closest to the upper right corner. In both data sets there are points with better specificity (more to the right in the QROC) and with better sensitivity (higher in the QROC). For example, in data set five if we chose to have sensitivity 1 (100%) the specificity is only 0.08 and the accuracy is 0.52. If specificity 1 is chosen, sensitivity becomes 0.04. In the optimal case for data set five, the values are sensitivity = 0.82, specificity = 0.88, and accuracy = 0.85.

Figure 2.2: ROC curves for two different methods on Down data.

2.3.3 Comparing areas under two ROC-curves

One might want to have an overall view of the performance of two different methods. To achieve this, it is possible to compare the area under the ROC curves or to compare the area under a part of the curves. Data set number five from the down data as described earlier will be used. In this case, two different methods are applied on the same data and want to use ROC analysis to see if there is any difference between the two methods. In Figure 2.2, the resulting ROC curves are shown. The dots are the results from the
preprocessing perceptron and the plus signs are from a fixed preprocessing perceptron. Both methods are described earlier in this chapter.

<table>
<thead>
<tr>
<th></th>
<th>PP</th>
<th>Fixed PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>0.8946</td>
<td>0.8691</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.0185</td>
<td>0.0210</td>
</tr>
</tbody>
</table>

Table 2.3: Area and standard error for the two ROC curves shown in Figure 2.2.

When the area beneath the curves are calculated (using Rockit [Rockit, 2002]), the results are as shown in Table 2.3. The correlation value between the two areas is 0.8882 and the area test statistic is $z = 2.6395$ and the two-tailed p-value is 0.0083. In other words, if the methods are compared at all cut-off points at once, the ordinary preprocessing perceptron has a significantly better performance than the fixed preprocessing perceptron on this data set.

Partial areas were compared, using the method described in [McClish, 1989], for specificity values 90-95%, which corresponds to the interval $[0.05, 0.1]$ on the x-axis, partial areas were also compared for sensitivity values 85-95%, which approximately corresponds to the interval $[0.25, 0.45]$ on the x-axis. In the interval, where the specificity is high there is a significant difference ($p = 0.02$) between the two. The PP area was 0.0322 and the fixed PP area was 0.0290. On the other hand, there was no significant difference in the interval where the sensitivity is high. The PP area was 0.1774 and the fixed PP area was 0.1733 ($p > 0.1$).

In the case of Down’s syndrome data, it is known that it is important to have a low false positive rate. If a decision had to be done about which one of these methods to use, the ordinary preprocessing perceptron should be chosen since its area is larger in this interval.

Rockit [Rockit, 2002] has been used in the experiments, which adapts data to a binormal model by using categorised data formed from ranks. According to [Hanley, 1996] the program is essentially semi-parametric, [...] it avoids the risk of biased, but possibly more precise, estimates of diagnostic accuracy estimates. They also show that it is justified to use his program even on data that does not look Gaussian. Another approach would be to use some of the distribution free methods mentioned earlier.
Chapter 3

Summary and Conclusions

An overview of the ROC curve analysis and its use within medicine and computer science has been given. The ROC analysis relies heavily on notations as sensitivity and specificity and these values depend on the specific data set. Even though the values in theory lie in the interval $[0,1]$, in practice the borders are decided by the data set. If QROC analysis is considered instead, which is built on two quality measures, a better way to compare different tests is given since all values have the same interval. In the beginning of this paper, terminology and notations from the field of medicine have been used. Since our use of ROC and QROC analysis are in the field of computer science or in the borderline medical informatics there are some differences in notation. Some of the differences are shown in Table 3.1.

<table>
<thead>
<tr>
<th>Medicine</th>
<th>Computer science/ medical informatics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usually compares different medical tests.</td>
<td>Usually compares different methods for decision support systems.</td>
</tr>
<tr>
<td>Use a gold standard (a test which is known to be “good”) and compare new tests with that.</td>
<td>Use the given diagnose in the data file as gold standard.</td>
</tr>
<tr>
<td>Wants the test to behave as close as possible to the diagnosis given by clinician.</td>
<td>Wants the method to behave as close as possible to the diagnosis given in the data file.</td>
</tr>
<tr>
<td>Use three different sampling methods depending on the risk or cost with testing and diagnosing</td>
<td>Use a data file gathered by other people (usually clinicians) and assumes that the sample is naturalistic, i.e. representative of the population.</td>
</tr>
<tr>
<td>Often includes cost or risk in assessing the tests, since they often are able to estimate these values for each test made.</td>
<td>Seldom includes cost or risk in their models. Are not able to assess these values in gathering data since they often get a data file without cost information.</td>
</tr>
</tbody>
</table>

Table 3.1: Differences in ROC terminology.

We have seen different ways to compare methods using ROC and/or QROC analysis. What method to use depends on the situation and the analysis might be different depending on if the distribution of the data is known or not. Both analysis methods that rely on binormal distributions and distribution free methods have been presented. A summary of
the most common ways to perform ROC/QROC analysis is:

**Point-wise comparison** Each point on the ROC/QROC curve corresponds to one test (one cut-off) and we get measurements of sensitivity and specificity values. We can get confidence regions for the values by the formula (if the sample is large enough)

\[ CR = X \pm t_{\alpha(2), n-1}SERR(X) \]

where \( X \) is the estimated sensitivity (or specificity). The standard error for the sensitivity is given by

\[ SERR(\hat{SE}) = \sqrt{SE(1-SE)/(N_0 \cdot P)} \]

and for specificity it is

\[ SERR(\hat{SP}) = \sqrt{SP(1-SP)/(N_0(1-P))}. \]

Sometimes it also interesting to compare two methods at a specific sensitivity or specificity, in that case, a \( \chi^2 \)-test can be used as described in the text.

**Sensitivity, specificity, efficiency and quality measures** The sensitivity and specificity values depend on the current situation, i.e., the level of the test, \( Q \). The level must be mentioned if sensitivity and specificity is used. It is better to use the quality measures

\[ \kappa(1, 0) = (SE - Q)/Q' \quad \text{and} \quad \kappa(0, 0) = (SP - Q')/Q \]

Efficiency is defined as \( EFF = TP + TN \) and have a corresponding quality measure

\[ \kappa(0.5, 0) = (PQ'+P0Q0\kappa(1, 0) + P0Q0\kappa(0, 0))/(PQ' + P0Q0). \]

**Optimal cut-off** The optimal cut-off depends on the situation. If the highest possible sensitivity is wanted, maximise \( \kappa(1, 0) \). When \( \kappa(1, 0) \) is maximised we have also maximised the negative predictive value. If we want to have the highest efficiency (the test and diagnose are the same), maximise \( \kappa(0.5, 0) \).

**Area under the curve** The larger the area under the curve is, the better discrimination performance has the test. The program Rockit is a good tool to use for calculating areas and comparing them.

**Partial area under the curve** Partial areas might be of interest sometimes. In some cases there might be no overall difference between methods while in one specific area there is one method that is significantly better than the other.

The use of ROC/QROC analysis is exemplified in the second part of the paper. Two data sets have been presented and four different ways to create decision support systems are shown. In summary, this paper has given an overview of what can be done with ROC/QROC analysis. The fact that it is important to use the right tool for the job at hand has also been stressed.
Bibliography


