# THE USE OF THE RECEIVER OPERATING CHARACTERISTIC IN THE RETRIEVAL OF DOCUMENTS IN A WEB DATA BASE

Carlos Bouza* and Marja Li**
*Facultad de Matemática y Computación, Universidad de La Habana
**Sofware Development Division, Institute of Computing Training.
bouza@matcom.uh.cu

**ABSTRACT**

Different methods are used for fitting the receiver operating characteristic curve of the similarity between the descriptors of the documents and the used in the search. They are compared using the search of scientific documents in a research institute. The accuracy of the procedures and the probabilities of error of each of them is computed.

**RESUMEN**

Diferentes mètodos son usados para ajustar el llamada 'receiver operating characteristic curve' de la similitud existente entre los descriptores de los documentos y los usados en su bùsqueda. Estos son comparados usando la búsqueda efectuada en de documentos científicos en una instituto de investigaci ò n. La exactitud de los procedimientos y las probabilidades de error de cada uno fue computada.

**Key Words**: empirical likelihood, kernel, area under the curve, dissimilarity, similarity

**MSC**: 62G07

## 1. INTRODUCTION

A broad collection of problems can be modeled by considering a signal game. A player sends a signal and the other matches it with his information. He gives what is supposed to be the best fitting. This is the current problem with information retrieval. It is commonly used in medical research for evaluating tests and procedures. The physician performs a series of tests and considers if the signal, send by them, are similar with the symptoms of a certain disease.

Actually we have this problem in the web when a user gives some key words and expect to obtain a document that fits with his needs. Our proposal is to use the receiving characteristic curve for evaluating the similarity between the documents in a data base and those that the user is looking for.

The paper is structured in 5 sections. The second section is devoted to discussing how receiver operating characteristic works. It can be considered as a general procedure for identifying signals,

as its name denotes. A discussion on signal decodification for assessing the efficacy of diagnostic tests in medicine is developed. It is used as a token for analyzing the role of ROC for assessing the discriminatory ability of tests. In the third section the connection between sensitivity and specificity is discussed and their role in determining the ROC curve. Section 4 present the methods that are generally used for estimating the area under the ROC and the curve fitting procedures. The use of empirical probability measures is the method commonly used. We propose the use of other methods which outputs allows to obtain a smooth curve. The method based on empirical likelihood is revised as well as the direct estimation of the distribution function through kernel based estimators. Section 5 presents a Game Theory signaling model and connect the decisions with a statistical game. Finally a comparison of the behavior of these models is made using real life data. The data was provided by the search of documents in a library through the search engine installed. The estimated area under ROC is computed for a set of estimation procedures and dissimilarity measures as well. Some clues for selecting the documents form the database are obtained from the experiments. The probability of rejecting a document, matching with the demand of the user, and of including un-matching document, is estimated for each combination of estimator-dissimilarity measure-area estimation method.

The results make a point on the use of this approach for interpreting signals in more general problems.

## 2. THE RECEIVER OPERATING CHARACTERISTIC

The development of Receiver Operating Characteristic (ROC) can be traced in Decision Theory. These curves assess the value of a diagnostic as they provide a standard measure of the ability of a test to correctly classify subjects. Its application in signal detection started with the study of radar images during World War II. The ROC curves enabled radar operators to distinguish between an enemy target, a friendly ship, or a simple noise. The problem was to decide what a blip on the screen represented. The ability of the operator (receiver) to identify the signals was called the Receiver Operating Characteristics. By the 1970's signal detection theory was considered as a useful tool for interpreting medical test results. The detection of cut-off points for radar equipment with different operators was considered a similar define cut-off points for diagnostic tests

The development of electronic communication poses similar problems. Among them are the use of signals send by a set of key information, as the key words in document, identification, for searching for documents; the decoding of images as in the handwritten problem, etc. These problems are actually very important. They have in common that some clues are given and the procedure must detect the matching. We can model this problem considering that a signal is obtained for analyzing a Data-Base and to determine which items identify that similar to their emission. Then we can model this problem using the framework of Games Theory where the player is looking for a winner decision rule.

The biomedical field uses signal decodification for assessing the efficacy of diagnostic tests in classifying subjects as healthy or not. Hence, ROC curves can assess the overall discriminatory ability of different potential indicator variables, characterizing the signal obtained. A measure for comparison should be defined for aiding in the selection of specific values of the variable. A

threshold is fixed for establishing a trade-off in the true positive rate and false positive rate. Values of the measure superior to the threshold parameter, determine that the signal is not from the objective population.

Medicine is clearly based on the need of developing adequate prognostics. Frequently, tests, drugs, or what we will call 'procedures', must be evaluated with respect to their ability to discriminate between those with and without a characteristic. Many Medical studies are based on continuous variables for developing predictions see for example Petrie-Sabin, C. (2000) and Zou et.al. (1997). A commonly used model used for evaluating the performance of diagnostic tests is the receiver operating characteristic (ROC) curve analysis.

The area under the curve described by ROC is often used to assess the discriminatory power of a procedure. It is also used for evaluating the accuracy of a diagnostic test. The expert would consider that the diagnostic ability, of a new procedure, is adequate only when a certain range of specificity is ensured. When a diagnostic test is based on a continuous measurement, a range of different decision thresholds or cutoff values may be investigated. The result permits to decide which value should be used to discriminate between patients, according to the outcome. Hence for measuring the accuracy of a continuous diagnostic test, the receiver operating characteristic (ROC) curve is often used. ROC curves are considered as a powerful and flexible tool to identify differentially expressed. The computation of the partial area under the curve gives information on its potentialities. Bamber (1975) and Liu et.al. (2005) show that the partial area under a ROC curve is equal to the probability of a constrained stochastic ordering. Lasko et.al. (2005) proved that each ROC character is isomorphic to a univariate probability distribution that possesses a moment generating function. These results give a sound basis to the theoretical potentialities of the decisions based on the analysis of ROC.

In any case from a mathematical point of view ROC summarizes the performance of a binary decision rule

**3. ON SPECIFITY AND SENSITIVIVITY**

The positive predictive value (PPV) of a test is the probability that a patient has a positive outcome, given that a procedure classifies him/her as positive. Its counterpart is the negative predictive value (NPV) which is the probability that a patient has a negative outcome, given that the outcome is negative. In medicine sensitivity is the probability that a patient has a positive test result given that the outcome is positive. Specificity is defined as the probability that a patient has a negative test result given that the result was negative.

The analysis of Sensitivity and Specificity is generally developed using the likelihood ratio

LR+= Senstivity(1-Specificity)-1.

It represents the increase in odds favoring the outcome given a positive test result.

We can define, the likelihood ratio of the probability of a negative test result when the outcome is positive to the probability of a negative test result if the outcome is negative, as

LR-= (1-Senstivity)(Specificity)-1.

$LR^-$ represents the increase in odds favoring the outcome given a negative test result.

The physicians consider as desirable to determine a test with high values for sensitivity and specificity.

The graph of sensitivity against (1 – specificity) is called the ROC curve. A test is evaluated as perfect if it has sensitivity=specificity=1. Sensitivity and specificity may not be regarded as equally important in the applications, because a false-negative finding may be more dangerous than a false-positive one or vice versa. Take for example to classify as a false negative a person with AIDS. It is not so important as to let undetected the sero-positiveness . Therefore in each investigation a cut-off should be chosen. In any case, if no judgment is made between the two the Youden's index is to be used.

J = sensitivity + specificity – 1

It provides a thumb rule for fixing an appropriate cut-off value. Note that $J \in [0, 1]$. It attains 1 for a perfect test and 0, if it has no diagnostic value.

Consider that we deal with a continuous variable such that, large values indicate an increased chance of a positive outcome. Then, when the cut-off value for this diagnostic variable is increased, the proportions of both true and false positives decrease. If the test is perfect then sensitivity=1 for any non-zero values of 1 – specificity. The ROC curve would start at the origin (0,0), goes vertically up the y-axis to (0,1) and then horizontally across to (1,1). A good test should have a behavior close to this ideal. Of course if a variable has no diagnostic capability, then a test based on that variable would be equally likely to produce a false positive or a true positive:

Sensitivity = 1 – specificity,

Or

Sensitivity + specificity = 1

This equality generates a diagonal line from (0,0) to (1,1) as the graph of the ROC curve.

Usually points are joined by straight lines but it is possible to fit a smooth curve. The use of a parametric model is the usual approach.

Sensitivity and specificity may not be invariant for a diagnostic test. They may depend on characteristics of the population under study. In medicine from ethical points of view go into the analysis. Then the ultimate decision on using or not a diagnostic test depends, not only on the ROC analysis but also on the ultimate benefit to the patient. The prevalence of the outcome, which is the pre-test probability, must also be known. Generally, there is a trade-off between sensitivity and specificity, and the practitioner must make a decision based on a compromise.

A high likelihood ratio for a positive result or a low likelihood ratio for a negative result indicates that the test is useful. As previously stated, a greater prevalence will raise the probability of a positive outcome given either a positive or a negative test result.

In applications we consider that the accuracy of the studied test depends on its capacity for separating the group being tested into two classes. One of them contains the ‘with the condition’

and the other the rest. From this point of view the area under ROC should be considered as a measure of discrimination, because we are interested in evaluating the ability of the test to correctly classify those with and without the characteristic.

The accuracy of a classification procedure may be measured by analyzing the area under the ROC curve. It is clear that an area of 1 represents a perfect test; an area of 0.5 represents a worthless test. A thumb rule is to classify the accuracy of a diagnostic test using the usual scholar classification and to consider that if the area is A:

- If $A \geq 0,9$ evaluate the test as excellent
- If $A \in ]0,80\ \ 0.90]$ evaluate the test as good
- If $A \in ]0,60\ \ 0.80]$ evaluate the test as fair
- If $A \in ]0,60\ \ 0,70]$ evaluate the test as poor
- If $A \leq 0,.50$ evaluate the test as bad

Usually a set of patients are correctly classified into two groups. Select at random a patient from the disease group and another from the no-disease group and consider that we apply the test on both patients. Take the rule:

"Classify the patient with the more abnormal test result in the disease group"

Then the area under the curve is the percentage of randomly drawn pairs for which the test correctly classifies the two patients in the random pair.

## 4. METHODS OF ESTIMATION OF THE AREA UNDER ROC

### 4.1 The Empirical Likelihood method of estimating ROC

A parametric approach uses maximum likelihood methods for fitting a smooth curve to the data points. Likelihood ratios permit to deal with tests with more than two possible results (not just positive/negative). The magnitude of it has an intuitive meaning in the evaluation of how strongly a given test's result influences in raising (rule-in) or diminishing lower (rule-out) the likelihood of the disease. Maximum likelihood estimation (MLE) method is used with frequency but it usually fails when we deal with small sample sizes.

Take X as the variable of the population of cases (with the characteristic) and Y of the population of controls. Denote by F the distribution of the cases. The value U=1-F(x) represents the proportion of false positives 1-E(U)=E(F(y))=P(Y>X)=d . Using this relationship the empirical likelihood (EL) theory is derived. Taking **p** = (p1, p2, … , pn) as the probability vector. We have the optimization problem

$$Ln(do)= Sup \{\Pi j=1npj \mid \sum j=1npj =1, \sum j=1npj (1-Uj -d0)=0\}$$

where $Uj = 1 - F(Yj),\ j = 1, 2, \ldots, n.$

The Uj's depend on the distribution function F. Hence using instead of F the empirical distribution Fn we may solve the estimation U*j =1-Fn(Yj) , 1,…,n. The new optimization problem is linked to the empirical likelihood and we have the new problem

L*n(do)= Sup {Πj=1npj ⏐ ∑ j=1npj =1, ∑ j=1npj (1-U*j -d0)=0}

The usual optimization procedure, through Lagrangian places to solve

Arg Ln(do)= {pj= n-1[1+λo(1-U*j-do)]-1, j=1,…,n ⏐ ∑j=1n (1-U*j-do)[1+λo(1-U*jdo)]-1=0}

As its maximum is n-n and it is obtained at pj =1/ n . The calculation of the value of λ0 is the numerical main difficulty in solving the optimization problem. The Likelihood ratio at d0 is equal to

R0=nΠj=1n pj=Πj=1n[1+λo(1-U*jdo)]-1

Then the empirical log-likelihood ratio is

.l(do)=-2log (Ro)=2∑j=1n [1+λo(1-U*jdo)]-

The standard theory of EL does not holds because the estimations U*j are non independent. Fortunately, see Theorem 1 of Qin-Zhou (2005*),* we can obtain a relatively simple solution. Take the joint sample of X and Y and order it. Using the common notation of rank based inferences Ri denotes the rank of X in the m cases and Si the rank of Y in the n controls. We may compute the rank means

$$\overline{R} = \frac{\sum_{i=1}^{m} R_i}{m}, \qquad \overline{S} = \frac{\sum_{i=1}^{n} S_i}{n},$$

and the variances

$$S_R^2 = \frac{\sum_{i=1}^{m} (R_i - i)^2 - m\left(\overline{R} - \frac{m+1}{2}\right)^2}{(m-1)n^2}, \qquad S_S^2 = \frac{\sum_{i=1}^{n} (S_i - i)^2 - n\left(\overline{S} - \frac{n+1}{2}\right)^2}{(n-1)m^2}$$

$$S^2 = \frac{mS_R^2 + nS_S^2}{m+n}$$

The asymptotic distribution of

$$\frac{m\sum_{j=1}^{n}(1-U^{*}_{j}-d_0)^2}{(m+1)nS^2}l(d_0)$$

is a $\chi 2(1)$.

Therefore hypothesis testing as well as interval estimation can be made

**4.2 Kernel estimation of ROC**

The empirical ROC curve is the most commonly used non-parametric estimator for the ROC curve. As it is an estimator of a probability distribution function the use of other estimators, different of the empirical measure, is natural. Lloyd (1998) proposed to use a kernel smoothing estimator for the ROC curve and derived that his proposal was more accurate in terms of the mean square error than the empirical ROC curve estimator. Kernel estimation of AROC is a challenge to usual techniques because it should be expected that kernel estimators should perform better than the usual estimator based on the frequencies

Consider that we deal with a real random variable X with a continuous distribution function F and a probability density function f. It is clear that the information on the probability measure provided by a sample (X1,…,Xn ) can be represented by a function K (x, Xi), centered in Xi, which obtains it maximum in the center of the region and decreases monotonically with the increase a function $\delta(x, Xi)$ . $\delta(\bullet)$ is a metric and K(•) is called kernel. An estimator of the density function f of X

$$f_h(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)$$

is the so called Parzen–Rosenblatt estimator of f(•). It depends on h=hn>0 which is a deterministic sequence known as band-width. It is assumed that it goes to zero when $n\to\infty$. Other theoretical derivations are a consequence of this estimator, see Silverman (1986) . Different types of kernels and their properties, the selection of an adequate bandwidth h etc, are part of the common work with this estimators The investigations on this subject establish that the kernel itself has not a serious influence in the performance of these estimators. The badndwidth hn, as the smoothing parameter, on the contrary plays the key role in the behavior of fh (x) because it determines the size of the x-centered neighborhoods. Note thath the graphic of fh(x) is a constant in each interval; hence it produces a stepwise figure. A sufficiently broad class of estimators is characterized by the following set of properties of the kerenel and the bandwidth:

a)$\int|K(z)|\,dz<\infty$

b) Sup $|K(z)| \to \infty$
c ) lim $z \to \infty$ $|zK(z)| = 0dz$
d) $K(z) \geq 0$
e) $\int K(z)dz = 1$.
f) $\int zK(z)dz = 0$
g) $\int z2\ K(z)dz < \infty$
h) $\int z2\ K(z)dz \neq 0$
i) lim $m \to \infty$ hmn =0
j) lim $z \to \infty$ mhmn $= \infty$

The kernel estimator of the distribution function at a point x is consequently

$$F_h(x) = \int_{-\infty}^{x} \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) dx = \int_{-\infty}^{x} f_h(t)dt$$

Using it we can estimate the distribution function of cases and controls and to use it for calculating the area under ROC.

Taking u as the standardized variable we can list some of the most popular kernel functions. See them in Table 4.1.

Table 4.1. Some kernels for density function estimation (u is an standardized variable and $\chi$( q) the characteristic function)

| Kernel | Formula |
|---|---|
| Cosines : K(u) | $25\pi\cos(\pi u/2)\ \chi(\ \|u\| \leq 1)$ |
| Epanechikov | $0{,}75(1-u2)\chi(\ \|u\| \leq 1)$ |
| Gaussian | $(2\pi)$-1exp{-u2/2} |
| Quartic | $15(1-u2)2\chi(\ \|u\| \leq 1)/16$ |
| Triangular: | $(1-\ \|u\|)\chi(\ \|u\| \leq 1)$ |
| Triweight | $35(1-u2)3\chi(\ \|u\| \leq 1)/32$ |
| Uniform | $0{,}5\chi(\ \|u\| \leq 1)$ |

**4.3 The most popular methods for estimating the AROC**

The usual method is to use the empirical distribution function for adjusting the curve.

The performance of a diagnostic variable is quantified by calculating the area under the ROC curve (AROC). The commonly used method relies on non-parametric theory. It is based on constructing trapezoids in a set of disjoint intervals under the curve. Each of them is an approximation of the area. Then the AROC is calculated by summing the areas of the trapeziums It is well known that, when sample observations are independent, the area under the receiver operating characteristic (ROC) curve corresponds to the Wilcoxon statistics if the area is calculated by the trapezoidal rule. Therefore another popular approach to estimate the AROC curve is to use a 'weighted' Wilcoxon-Mann-Whitney statistic considering its desirable statistical properties. Then a non parametric testing procedure allows comparing the partial area under two receiver ROC curves. Specialized statistical packages allow developing the calculations involved

**4.4 The Comparisson of two ROC's**

The ability of two continuous variables to diagnose an outcome can be compared using ROC curves and their AROC's. A formal test should be used for establishing whether the difference between them is significant. The difference in the shape of the curves may be important because it can establish if for very low levels of sensitivity correspond high levels of specificity. If a cut-off is selected for a high level of specificity, the real discriminating power can be differentiated.

Nonparametric comparison of areas under correlated ROC curves. Provides point and confidence interval estimates of each curve's area and of the pairwise differences among the areas. Tests of the pairwise differences are also given. Any contrast among the areas may be estimated and tested

**5. IDENTIFICATION AS A STATISTICAL GAME**

Let us consider that an identification system proceeds as follows

1. An information is obtained through a variable X
2. The identifier $\psi$ is a system that transforms X and compares it with the Data-base.
3. A set C(+) of possible candidates identified as plausible (templates) is determined
4. The items in C(+) are listed and a more sophisticated procedure identifies correctly them.

A pay-off is established for evaluating how effectively the system has conformed C(+). The player evaluates the consistency of the system generated decisions by rating the bad identifications (False Matching Rate, FMR) and the incorrect rejects (False Non Matching Rate,

FNMR). These events have probabilities

P(+)=P(FMR)= probability of identifying incorrectly an item as consistent (matching)

P(-)=P(FNMR)= probability of identifying incorrectly an item as no consistent (non-matching).

Then, we have again the problem of the false negative. An error is present if the test procedure reports incorrectly that the quality is not present while it really is present

Measuring the similarity between documents and queries has been extensively studied in information retrieval See Meltzer et.al. (2005), The detection algorithms are unable to be correct always. That is the case when the user gives a set of key words and the search procedure gives a document that is not matching with the subject. The proportion of them among the observed documents is the false negatives rate. Again the sensitivity and specifity ideas appear because the player is interested in knowing

Rate of false negatives = number of false negatives / number of de positives

Rate of false positives = number of false positives / number of negatives

This fraction is called test size.

Using $\psi(X)$ a score $\sigma(\psi(X), \psi(X\ k))$ is evaluated in the observed item, k=1,...,n. This similarity measures generally takes value in the interval [0, 1]. Hence we accept that k matched with the signal when $\sigma(\psi(X), \psi(Xk\ ))\cong 1$ . The problem to be solved is to fix a threshold parameter 0<T<1 such that k is considered as a match when it belongs to

$C(+)=\{k|\sigma(\psi(X), \psi(X\ k)) \geq T\}$

With a sample s of size n, we obtain $\{\psi(X1\ ),..., \psi(X\ M)\}$. The player looks for a decision rule such that for a new item '0" he can decides if $k(0)\in C(+)$ or not. Using the previously developed ideas a rule is modeled by the statistical game:

Accept H(A) if $\sigma(\psi(Xk(0)\ ), \psi(X\ k))\geq T$

Accept H(R) if $\sigma(\psi(Xk(0)\ ), \psi(X\ k))<T$

Accepting H(A) is related with the possibility of failing (error of type I, a false positive). To accept H(R) is also related with the possibility of failing (error of type II, false negative). Note that

FMR=Prob. $(H(A)|H(R)\ ))=\alpha$

FNMR=Prob. $(H(R)|H(A)\ ))=\beta$

Generally with do not know the real distribution $P(\psi(Xk(0))|H(A))$ of identifying the matching items (genuine distribution). The distribution $P(\psi(Xk(0))|H(R))$ is also unknown ( impostor distribution). The statistical game assumes that the player is able to calculate the probabilities of deciding incorrectly:

$\alpha=\int 0T\ d\ P(\psi(Xk(0))|H(A))$

$\beta=\int T1\ d\ P(\psi(Xk(0))|H(\ R))$

The information from the nature is obtained by sampling. As the distributions are unknown the sample s is divided into two sub-samples , s(1) and s(2), of size n(1) and n(2). The first one is used for training the system and the second one for evaluating the effectivity of the signal identification system. Denote by $\{S(k)=\psi(Xk), k\in s(1)\}$ the set of signals obtained from the training sub- sample . It allows estimating the genuine distribution. Evaluating the performance of the system we can estimate the impostor distribution.

We have that the probability of observing a false positive in the search is

$\alpha$=1-(1-FMR)n(1)

Considering that FMR$\cong$0 a Taylor Series expansion can be developed with the second term. We obtain that

$\alpha\cong$n(1)-1FMR

Hence using s(2) we can estimate it by

FMR*= number of incorrectly identified /n(2)

Note that using the n(1) evaluations with the items not included in the data-base, the expression of the probability remains the same and we have that the probability of obtaining a false negative is FNMR=$\beta$. Therefore it is estimated by

FNMR*= number of bad identified item in s(2) /n(2).

T is obtained by fixing a value of $\alpha$ and determining T$\alpha$. or by projecting the intersection point of the estimated impostor and genuine curves.

## 6. A STUDY OF THE PERFORMANCE OF A SEARCH ENGINE

We will deal with the collection of documents matching with the signal given by key words for a fixed set of N descriptors. The signal is a binary valued vector X=(X1 ,..,XN ) with Xj$\in${0,1}. Xi =1 if the descriptor is present. Once a sample is evaluated we compute

S11=number of matchings

S10=number of words used by the user matching with the document but not present in it.

S01=number of words in the document matching with the theme but not reported by the user.
S00=number of unmatched words evaluated as un-matching.

We have the following similarity and dissimilarity measured for this problem

Table 5.1. Similarity and Dissimilarity measures for binary vectors

| Name of the measure | Similarity | Dissimilarity |
|---|---|---|
| Jaccard-Needham | S11(S11+S10+S01)-1 | (S10+ S10 )(S11+S10+S01)-1 |
| Dice | S11(2S11+S10+S01)-1 | (S10+ S10 )(2S11+S10+S01)-1 |
| Correlation | (S10S00–S01S10 )(S11S00+S10+S01)-1 | 1-(S10S00–S01S10 )(S11S00+S10+S01)-1 |
| Yule | (S10S00–S01S10 )(S11S00+S10+S01)-1 | (S10S01)(S11S00+S10+S01)-1 |
| Russell-Rao | S11N-1 | (n-S11 )N-1 |
| Sokal-Michener | (S11+S00)N-1 | (2S10+2S01 )(S11+S00+2S10+2S01)-1 |
| Rogers-Tamamoto | (S11+S00 )(S11+S00+2S10+2S01)-1 | (2S11+2S00 )(S11+S00+2S10+2S01)-1 |
| Kulzinsky | S11(S10+S01)-1 | (N+S11+S01 +S01)(N+S10+S01)-1 |

Nuñez et.al. (2003) developed a study of similarity measures in the evaluation of environmental systems reasoning. Sahami and Heilman (2006) considered the use of kernels for evaluating the similarity of texts. We will analyze the behavior of these similarities measures for evaluating the effectivity of the search engine of a library. The library contents more than 31 000 documents. It is located at an institute with around 2 000 researchers and graduate students.

An experience was developed using the web connections with the library. The users were invited to report if the provided documents matched with their search or not. The number or daily reports were between 67 and 249. This information was used for computing each of the dissimilarity measures in Table 5.1. The dissimilarity was computed using the words in the title and key words of the document and the descriptor given by the user.

The AROC was determined using the measurements made during 320 days. The study of the distribution of each measure should provide the ROC. Then accuracy of the measure for describing the behavior of the identification was made. The area under the ROC curve was estimated using different criteria. The empirical ROC was computed using. StatsDirect which calculates the area under the curve directly by an extended trapezoidal rule, see Press et al. (1992) and by a non-parametric method analogous to the Wilcoxon-Mann-Whitney test, see Hanley and McNeil (1982). The empirical likelihood estimates was computed using a MatLab code developed by the authors. The estimation of the density distributions was developed using the package SAS/STAT which provides procedures for nonparametric density estimation using the method of kernel density estimation. The selection of the bandwidth was made using the automatic bandwidth selection proposed by Silverman (1986).

The computed AROC is given in table 5.2.

Table 5.2, The estimated AROC using the Trapezoidal method and dissimilarity measures

| Estimator of AROC | Jaccard -Needham | Dice | Correlation | Yule | Russell -Rao | Sokal -Michener | Kulzinsky | Rogers -Tamamoto |
|---|---|---|---|---|---|---|---|---|
| Empirical likelihood | 0.67877 | 0.86556 | 0.68453 | 0.83414 | 0.82161 | 0.88297 | 0.70801 | 0.57899 |
| Cosines | 0.46556 | 0.83487 | 0.49536 | 0.91853 | 0.85951 | 0.87977 | 0.42548 | 0.49769 |
| Epanechikov | 0.78093 | 0.85335 | 0.83595 | 0.91225 | 0.83369 | 0.88825 | 0.55406 | 0.54259 |
| Gaussian | 0.86051 | 0.86418 | 0.76777 | 0.85837 | 0.90773 | 0.87086 | 0.74363 | 0.86441 |
| Quartic | 0.7966 | 0.87174 | 0.78469 | 0.80049 | 0.86777 | 0.82548 | 0.59768 | 0.89843 |
| Triangular | 0.86777 | **0.90647** | 0.74025 | 0.87836 | 0.93516 | 0.89688 | **0.88711** | 0.86777 |
| Tri-weight | **0.92368** | 0.90508 | **0.96894** | **0.93388** | **0.93696** | **0.8994**1 | 0.84325 | **0.96777** |
| Uniform | 0.38633 | 0.77899 | 0.34717 | 0.70381 | 0.68711 | 0.61514 | 0.30801 | 0.38426 |

An analysis of Table 5.2 suggests that the use of the Triweight kernel is the best alternative for estimating the AROC. The measures of Dice, Yule, Russel-Rao and Sokal-Michener have similarly good coverage in al the cases.

The results in Table 5.3 give a similar evaluation of the behavior of the measures. The performance of the curve fitting model is differentiated. Triweight is not longer the best method. The Gaussian and the Triangular kernels produce the best fitting for three dissimilarity measures each one.

The Uniform kernel produces the worst estimators in both cases.

Table 5.3, The AROC using the Weighted Wilcoxon-Mann-Whitney statistic method and dissimilarity measures

| Estimator of AROC | Jaccard -Needham | Dice | Correlation | Yule | Russell -Rao | Sokal -Michener | Kulzinsky | Rogers -Tamamoto |
|---|---|---|---|---|---|---|---|---|
| Empirical likelihood | 0.86913 | 0.83598 | 0.88426 | 0.88369 | 0.82161 | 0.83414 | 0.46556 | 0.78093 |
| Cosines | 0.84841 | 0.87114 | 0.80404 | 0.89932 | 0.85951 | **0.91853** | 0.83487 | 0.86051 |
| Epanechikov | 0.85129 | 0.85952 | 0.86971 | 0.86755 | 0.73369 | 0.91225 | 0.65335 | 0.7966 |
| Gaussian | **0.94259** | **0.92637** | **0.90404** | 0.87836 | 0.90773 | 0.85837 | 0.86418 | 0.86777 |
| Quartic | 0.87916 | 0.84133 | 0.88195 | **0.91633** | 0.86777 | 0.80049 | 0.91174 | 0.92368 |
| Triangular | 0.90617 | 0.91422 | 0.84025 | 0.87836 | **0.93516** | 0.89688 | **0.97450** | **0.96777** |
| Tri-weight | 0.86913 | 0.83598 | 0.89124 | 0.88173 | 0.81596 | 0.89941 | 0.84325 | 0.83921 |
| Uniform | 0.58540 | 0.67899 | 0.54717 | 0.50381 | 0.38711 | 0.61514 | 0.30801 | 0.38426 |

A threshold was determined fixing that a document with a dissimilarity larger than 0,6 is considered as not-matching. An evaluation sample was collected during a month. FMR* and FNMR* were computed. The results appear in Tables 5.4 and 5.5. The best global performance is obtained by the Triweight kernel in terms of the probilitly of the first type of error. The recommendation is to use Jaccard-Needham measure and the Triweight kernel when the trapezoidal s used.

Table 5.4. The Empirical Probability Errors of the AROC using the using the Trapezoidal method and dissimilarity measures

| Estimator of AROC | | Jaccard -Needham | Dice | Correlation | Yule | Russell -Rao | Sokal -Michener | Kulzinsky | Rogers -Tamamoto |
|---|---|---|---|---|---|---|---|---|---|
| Uniform | $\alpha^*$ | 0,229 | 0,271 | 0,286 | 0,260 | 0,236 | 0,296 | 0,229 | 0,271 |
| | $\beta^*$ | 0,559 | 0,675 | 0,699 | 0,493 | 0,230 | 0,389 | 0,589 | 0,675 |
| Empirical likelihood | $\alpha^*$ | 0,176 | 0,167 | 0,142 | 0,161 | 0,246 | 0,290 | 0,174 | 0,167 |
| | $\beta^*$ | 0,813 | 0,885 | 0,457 | 0,827 | 0,820 | 0,412 | 0,551 | 0,835 |
| Cosines | $\alpha^*$ | 0,182 | 0,231 | 0,258 | 0,147 | 0,278 | 0,134 | 0,182 | 0,231 |
| | $\beta^*$ | 0,419 | 0,667 | 0,346 | 0,564 | 0,311 | 0,387 | 0,419 | 0,567 |
| Epanechikov | $\alpha^*$ | 0,101 | 0,130 | 0,119 | 0,152 | 0,163 | 0,148 | **0,101** | 0,160 |
| | $\beta^*$ | 0,352 | 0,420 | 0,366 | 0,480 | 0,235 | 0,422 | 0,223 | 0,420 |
| Gaussian | $\alpha^*$ | 0,116 | **0,114** | 0,099 | 0,159 | 0,169 | **0,113** | 0,116 | 0,114 |
| | $\beta^*$ | 0,485 | **0,141** | **0,135** | 0,110 | **0,113** | 0,157 | **0,185** | 0,241 |
| Quartic | $\alpha^*$ | 0,208 | 0,257 | 0,228 | 0,172 | 0,358 | 0,211 | 0,208 | 0,257 |
| | $\beta^*$ | 0,431 | 0,139 | 0,155 | 0,314 | 0,182 | **0,118** | 0,188 | 0,269 |
| Triangular | $\alpha^*$ | 0,248 | 0,185 | 0,157 | 0,174 | 0,235 | 0,157 | 0,248 | 0,185 |
| | $\beta^*$ | 0,569 | 0,488 | 0,534 | 0,392 | 0,420 | 0,373 | 0,492 | 0,588 |
| Tri-weight | $\alpha^*$ | **0,09** | 0,115 | **0,105** | **0,095** | **0,139** | 0,119 | 0,109 | **0,159** |

The results of the analysis of Table 5.6 are close to those in Table 5.5. The triweight is the best overall kernel procedure but Dice and the Gaussian kernel is the best choice for selecting the documents.

| Estimator of AROC | | Jaccard -Needham | Dice | Correlation | Yule | Russell -Rao | Sokal -Michener | Kulzinsky | Rogers -Tamamoto |
|---|---|---|---|---|---|---|---|---|---|
| Empirical likelihood | $\alpha^*$ | 0,154 | 0,184 | 0,140 | 0,184 | 0,252 | 0,222 | 0,149 | 0,155 |
| | $\beta^*$ | 0,522 | 0,598 | 0,713 | 0,495 | 0,219 | 0,392 | 0,522 | 0,685 |
| Cosines | $\alpha^*$ | 0,194 | 0,251 | 0,248 | 0,134 | 0,291 | 0,111 | 0,198 | 0,251 |
| | $\beta^*$ | 0,395 | 0,705 | 0,349 | 0,571 | 0,328 | 0,393 | 0,402 | 0,526 |
| Epanechikov | $\alpha^*$ | 0,121 | 0,124 | 0,122 | 0,155 | 0,143 | 0,158 | **0,097** | 0,165 |
| | $\beta^*$ | 0,334 | 0,417 | 0,367 | 0,435 | 0,196 | 0,435 | 0,215 | 0,464 |
| Gaussian | $\alpha^*$ | 0,117 | **0,088** | 0,101 | 0,143 | 0,173 | **0,091** | 0,156 | 0,244 |
| | $\beta^*$ | 0,477 | **0,185** | **0,144** | 0,108 | **0,105** | 0,130 | **0,105** | 0,254 |
| Quartic | $\alpha^*$ | 0,212 | 0,238 | 0,231 | 0,134 | 0,343 | 0,232 | 0,301 | 0,265 |
| | $\beta^*$ | 0,332 | 0,202 | 0,150 | 0,327 | 0,310 | **0,107** | 0,190 | 0,261 |
| Triangular | $\alpha^*$ | 0,211 | 0,179 | 0,118 | 0,167 | 0,365 | 0,102 | 0,232 | 0,173 |
| | $\beta^*$ | 0,524 | 0,457 | 0,537 | 0,388 | 0,411 | 0,327 | 0,463 | 0,542 |
| Tri-weight | $\alpha^*$ | **0,115** | 0,118 | **0,105** | **0,091** | **0,108** | 0,124 | 0,112 | **0,115** |
| | $\beta^*$ | **0,216** | 0,213 | 0,244 | **0,213** | 0,232 | 0,247 | 0,218 | **0,206** |
| Uniform | $\alpha^*$ | 0,237 | 0,278 | 0,185 | 0,259 | 0,218 | 0,302 | 0,231 | 0,203 |
| | $\beta^*$ | 0,792 | 0,881 | 0,398 | 0,745 | 0,727 | 0,425 | 0,493 | 0,773 |

## REFERENCES


Bewick V, Cheek L, Ball J. (2004): Qualitative data – tests of association. **Crit Care.** 8,46–53.

Campbell, M. J.; Machin, D. (1999): **Medical Statistics: A Commonsense Approach.** Wiley; Chichester,

Hanley J. A, McNeil BJ. (1983): A method of comparing the areas under receiver operating characteristic curves derived from the same cases. **Radiology. .** 148, 839–843.

Horová, I., Forbelská, M. AND Zelinka, Jiří. (2006): Comparative Study of the Estimation of the Area Under the ROC Curves. In Proceedings of the International Workshop on Spatio-Temporal Modelling (METMA3).177-180. **Instituto de Estadistica de Navarra,** Pamplona.

Hsieh. F. and Turnbull, B. W. (1996): Non-parametric and semi-parametric estimation of the receiver operation characteristic curve. **Ann. Statist.** 24, 25–40.

Lasko, T. A., Bhagwat, J. G.; Zou, K.H. and Ohno-Machado, L. (2005): The use of receiver operating characteristic curves in biomedical informatics. **Journal of Biomedical Informatics,** 38, 404-415.

Liu, A.; E. F. Schisterman and C. Wu (2005): Nonparametric Estimation and Hypothesis Testing on the Partial Area Under Receiver Operating Characteristic Curves Communications in Statistics –Theory and Methods, 34, 2077-2088

Lloyd . C.J. (1998): The use of smoothed ROC curves to summarise and compare diagnostic systems. **J. Amer. Statist. Assoc.** 93, 1356–1364.

Lloyd C. and Z. Yong (1999): Kernel estimators of the ROC curve are better than empirical. **Statist. Probab. Lett.** 44, 221–228.

Margolis D.J; Bilker W.; Boston R.; Localio R. and Berlin J.A. (2002) : Statistical characteristics of area under the receiver operating characteristic curve for a simple prognostic model using traditional and bootstrapped approaches . **Journal of Clinical Epidemiology**, 55, 518-524.

Mei-Ling Ting Lee, Bernard A. Rosner (2001): The Average Area under Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach Based on Generalized Two-Sample Wilcoxon Statistics. **Applied Statistics,** 50,337-344.

Metzler, D., Bernstein, Y., Croft, W.B., Moffat, A., and Zobel, J. (2005): Similarity measures for tracking information flow. In **Proceedings of ACM Conference on Information and Knowledge Management, 517-524.. ACM** New York,

Nùñez , H., Sànchez-Marrè, M. Cortè, U. and Comas, J. (2003): A comparative study on the use of similarity measures in case based reasoning to improve the classification of environmental system situations. **Environmental Modelling & Software** 20, 3-14.

Petrie, A.; Sabin, C. (2000): **Medical Statistics at a Glance.** Blackwell. Oxford.

Whitley E, Ball J. Statistics (2002): Nonparametric methods. **Crit Care.** 2, 509–513.

Press, W. H.S.A. Teukolsky, W.T. Vetterling and B.P. Flannery (1992):N**umerical Recipes in C: the art of scientific computing,** Second Edition, Cambridge University Pres, Cambridge.

Sahami, M. and Heilman, T.(2006): A web-based kernel function for measuring the similarity of short text snippets. In **Proceedings of WWW 377-386, 2006. ACM**, N. York.

Tubbs. J. D. (1989): A note on binary template matching. **Pattern ecognition,** 22, 359–365.

Zhang, B. and S. N. Srihari. (2003): Binary vector dissimilarities for handwriting identification. In **Proceedings SPIE, Document Recognition and Retrieval X, 155–166, SPI,** Santa Clara..

Zou, K.H. , W.J. Hall and D.E. Shapiro (1997), Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. **Statist. Med. 16** (1997), pp. 2143–2156.