On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy

Markus Junker
German Research Center for
Artificial Intelligence (DFKI) GmbH
P.O. Box 2080, D-67608 Kaiserslautern, Germany
markus.junker@dfki.de

Rainer Hoch
SAP AG, Basis Systems & Services
P.O. Box 1461, D-69190 Walldorf, Germany
rainer.hoch@sap-ag.de

Andreas Dengel
German Research Center for Artificial Intelligence (DFKI) GmbH
P.O. Box 2080, D-67608 Kaiserslautern, Germany
andreas.dengel@dfki.de

Abstract

In document analysis, it is common to prove the usefulness of a component by an experimental evaluation. By applying the respective algorithms to a test sample, some effectiveness measures such as recall, precision, and accuracy are computed. The goal of such an evaluation is two-fold: on the one hand it shows that the absolute effectiveness of the algorithm is acceptable for practical use. On the other hand, the evaluation can prove that the algorithm has a better or worse effectiveness than another algorithm. In this paper we argue that the experimental evaluation on relative small test sets – as is very common in document analysis – has to be taken with extreme care from a statistical point of view. In fact, it is surprising how weak statements derived from such evaluations are.

1 Introduction

The task of document analysis is to transform printed documents into an equivalent electronic representation. Typical problems of document analysis systems are: image processing, layout segmentation, structure recognition, optical character recognition (OCR), but also the selective extraction of information such as the type of document [3, 1]. An important subtask of information extraction that is becoming more and more important is document categorization, for example. By document categorization we mean the automatic classification of documents into different types allowing workflow management, the automatic routing or archiving of documents, the search for notes in online service systems to solve customer requests, and many more practical applications. A central question is how to evaluate the effectiveness of such complex document analysis systems involving rather distinct components. In document categorization, e.g., we are interested in stating something about the performance of a particular algorithm. In addition, we are interested in comparing different document categorization algorithms by their effectiveness [8, 11, 6].

The effectiveness of text retrieval systems is usually given by the well-known IR standard measures *recall* and *precision* [10]. The *recall* of a text retrieval system can be defined as the ratio of the number of relevant documents returned to the total number of relevant documents for the user query in the collection. The *precision* is the ratio of the number of relevant documents returned to the total numbers of documents for a given user query. In Machine Learning, the *accuracy* is a widespread effectiveness measure to evaluate a classifier's performance [9]. It is defined as the ratio of the number of correctly classified items to the total number of items. The error rate, defined as $(1 - \textit{accuracy})$, is also very common in machine learning.

In this paper, we propose that the standard measures recall, precision, and accuracy/error rate are adequate measures to compute the effectiveness of document analysis components. Table 1 defines these standard measures in the context of different document analysis components. Due to the simple relation between error rate and accuracy, the error rate is not listed in the table.

It is important to note that these measures are all computed on a randomly chosen test set. Thus, they only provide the effectiveness of some algorithms on this particular set. It is unclear whether the measured effectiveness will hold

character/word recognition	
recall	# (character/word x correctly recognized) / #(character/word x occurs)
precision	# (character/word x correctly recognized) / #(character/word x recognized)
accuracy	# (characters/words correctly recognized) / # (all characters/words)
structure recognition	
recall	# (label x correctly assigned to some segment) / # (segment with label x occurs)
precision	# (label x correctly assigned to some segment) / # (segment with label x recognized)
accuracy	# (labels correctly assigned) / # (total number of label assignments)
document categorization	
recall	# (document correctly assigned to category x) / # (documents belonging to category x)
precision	# (document correctly assigned to category x) / # (documents assigned to category x)
accuracy	# (documents correctly categorized) / # (documents)
information extraction	
recall	# (information of type x correctly extracted) / # (information of type x occurs)
precision	# (information units of type x correctly extracted) / # (information units of type x recognized)
accuracy	# (information units correctly extracted) / # (information units to extract)

Table 1. Recall, precision, and accuracy with respect to some document analysis components

for new documents from the same domain. In particular, there is some risk that the algorithm will perform as well on new documents. This is a problem when we would like to guarantee for the effectiveness of a particular algorithm in the application. Another problem arises when we want to compare the effectiveness of two algorithms. Even if algorithm 1 outperforms algorithm 2 on a test set, this may not be a clear indication that 1 is really better than algorithm 2.

A closer look at the ratios given in Table 1 helps to gain a statistical access to the problems. It reveals that all ratios computed there have the form

$$\frac{\#(A \cap B \text{ occurs in test set})}{\#(B \text{ occurs in test set})}$$

with A and B being some events. In document categorization, e.g., the recall is computed by dividing the number of documents that are assigned to category x and belong to category x, by the number of documents which belong to category x.

Using the terminology of events A and B, a more interesting question is whether in general if B holds, A also holds. This can be expressed by the conditional probability $p = prob(A|B)$. The ratio we compute on the test set then is the maximum likelihood estimate for the unknown probability:

$$p = prob(A|B) \approx \frac{\#(A \cap B \text{ occurs in test set})}{\#(B \text{ occurs in test set})}$$

In the following sections, we rely on this maximum likelihood estimate to analyse the effectiveness of document analysis components in more detail.

2 Lower bounds for the effectiveness of a component

In many practical applications a lower bound for the component's real effectiveness must be provided. Having only small test sets, there is a high risk that the real effectiveness is much smaller than the maximum likelihood estimate. This poses the question, which minimal effectiveness of a classifier can be guaranteed by the evaluation on the test set with some error probability. In statistical terms the problem can be formulated as follows: Given a maximum likelihood estimate $\frac{x}{n}$ for an unknown probability $p = prob(A|B)$, how can we compute a confidence interval which contains p with a low error probability? Computing the estimate $\frac{x}{n}$ can be seen as making n times a Bernoulli experiment with the unknown probability p. In this interpretation, the value x denotes the number of successes (i.e. the number of times A holds).

The literature provides standard techniques to compute an interval $[p_{min}, p_{max}]$, which contains p with an error probability of γ. In our case, we are interested in intervals of the form $[p_{min}, 1]$, i.e. we need a lower bound p_{min} for p. In the following we describe an estimation of p_{min} which is adapted from [7].

Let X be a random variable which describes the number of successes in n experiments. The probability of having exactly k or less than k successes is computed using p as:

$$prob(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

First, we search for:

$$x_{min} \text{ with } prob(X < x_{min}) = \gamma$$

The value x_{min} can be approximated as follows:

$$x_{min} = \max_x \left\{ \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i} \leq \gamma \right\} \quad (*)$$

We can now show the following equivalence (compare [7], 173pp):

$$x < x_{min} \iff p \leq p_{min}$$

Using this, we get

$$prob(P \leq p_{min}) = prob(X < x_{min})$$

with P being a random variable for the real probability. For an observed x, the value of p_{min} can be computed approximately using formula (*).

Figure 1 shows the value of p_{min} based on the maximum likelihood estimate $\frac{x}{n}$ for different sizes n and the error probability $\gamma \leq 0.05$, Using the figure it is easy to find the lower bound p_{min} based on an estimate $\frac{x}{n}$. If, e.g., the maximum likelihood estimate for p is 0.5 with $n = 20$, the real value of p is bigger than 0.3 with an error probability of $\gamma = 0.05$. If the maximum likelihood estimate is 1.0 with $n = 20$ we can guarantee a lower bound of the component's effectiveness of just 0.86 with $\gamma = 0.05$. The same estimate 1.0 with $n = 100$ increases the lower bound to 0.97. The examples illustrate that for smaller sizes of n the lower bound is quite a bit lower than the estimate.

3 Comparing components

We now turn to the problem of comparing two classifiers by their maximum likelihood estimates on a test set. The interesting question here is whether a measured difference in the estimates $\frac{x_1}{n_1}$ and $\frac{x_2}{n_2}$ for p_1 and p_2 is significant or not. The probability of succeeding exactly x_1 times in n_1 experiments and x_2 times in another n_2 experiments can be computed using the unknown probabilities p_1 and p_2:

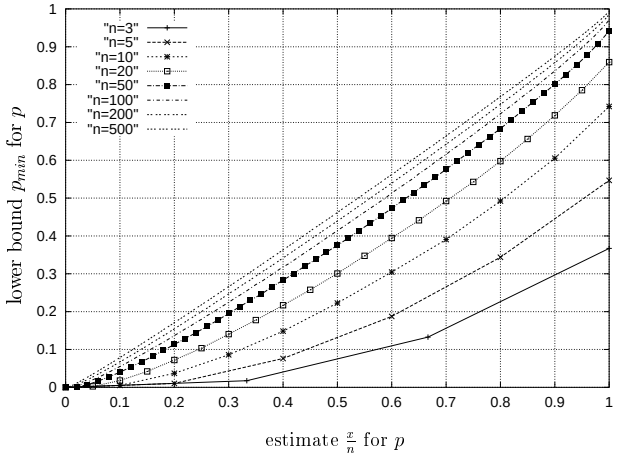

Figure 1. Lower bound for p based on estimate $\frac{x}{n}$ (error probability $\gamma \leq 5\%$)

$$\binom{n_1}{x_1}\binom{n_2}{x_2} p_1^{x_1}(1-p_1)^{n_1-x_1} p_2^{x_2}(1-p_2)^{n_2-x_2}$$

If the hypothesis $p_1 = p_2$ holds this can be simplified to:

$$f(x_1, x_2) = \binom{n_1}{x_1}\binom{n_2}{x_2} p^{x_1+x_2}(1-p)^{n_1+n_2-(x_1+x_2)}$$

Given $x_1 + x_2 = s$, the conditional common probability of x_1 and x_2 does not depend on p. It can be computed using the random variable S introduced by s:

$$g(x_1, x_2|s) = \frac{f(x_1, x_2)}{prob(S = s)}, (x_1 + x_2 = s)$$

Since S is normally distributed with the parameters $n_1 + n_2$ and p under the hypothesis $p = p_1 = p_2$ we get

$$prob(S = s) = \binom{n_1 + n_2}{s} p^s (1-p)^{n_1+n_2-s}$$

with

$$g(x_1, x_2|s) = \frac{\binom{n_1}{x_1}\binom{n_2}{x_2}}{\binom{n_1+n_2}{s}}, (x_1 + x_2 = s)$$

And since $x_2 = s - x_1$ holds:

$$h(x_1|s) = \frac{\binom{n_1}{x_1}\binom{n_2}{s-x_1}}{\binom{n_1+n_2}{s}}, x_1 = 0, 1, \ldots, min(n_1, s)$$

Using the above formula we can compute a confidence interval $[x_{1,min}, \ldots, x_{1,max}]$ which contains x_1 with an error probability of γ in the case of $p_1 = p_2$. The interval boundings should satisfy the conditions $prob(X_1 < x_{min}) \leq \frac{\gamma}{2}$ and $prob(X_1 > x_{max}) \leq \frac{\gamma}{2}$

$$x_{1,min} = \max_k \left\{ \sum_{i=0}^{k} h(i|s) \leq \frac{\gamma}{2} \right\} \quad (1)$$

$$x_{1,max} = \max_k \left\{ \sum_{i=s-k}^{s} h(i|s) \leq \frac{\gamma}{2} \right\} \quad (2)$$

If we observe that x_1 is outside the computed interval $[x_{1,min}, \ldots, x_{1,max}]$, the hypothesis $p_1 = p_2$ will be rejected with error probability γ.

The derived significance test can be applied to the initial problem in the following way. Using $s = x_1 + x_2$ and our chosen error probability γ, $x_{1,min}$ and $x_{1,max}$ can be computed by the formulas (1) and (2). The measured difference is significant, if $x_1 \leq x_{1,min}$ or $x_1 \geq x_{1,max}$ holds (respectively $\frac{x_1}{n} \leq \frac{x_{1,min}}{n}$ or $\frac{x_1}{n} \geq \frac{x_{1,max}}{n}$).

Figure 2 shows the minimal increase in the maximum likelihood estimates required for different values $n = n_1 = n_2$ and $\gamma = 0.05$. By the choice of n_1 and n_2 we can only get certain discrete values for the maximum likelihood estimates. For easier readability we did not mark these values for $n \geq 100$. Using Figure 2 we can state that with $n_1 = n_2 = 50$ an increase in the maximum likelihood estimate from 0.5 to 0.7 for the effectiveness of a component is not significant with $\gamma = 0.05$. On the other hand, an increase by the same amount from 0.8 to 1.0 is significant with $\gamma = 0.05$. With an estimate of 0.9 even an increase to the maximum value of 1.0 is not significant anymore (at least with $\gamma = 0.05$).

Since in the case of precision $n_1 = n_2$ does not hold in general, the significance of an increase in precision cannot directly be derived from Figure 2. A pessimistic decision can be made by choosing $n = min\{n_1, n_2\}$, as the following example illustrates: Let us assume we measure a precision of 0.8 for a component with $n_1 = 50$ and a precision of 1.0 for another component with $n_2 = 100$. Using Figure 2 ($\gamma = 0.05$) we can see that for $n = min\{50, 100\} = 50$, already an increase from 0.8 by 0.16 is significant. Using the exact values n_1 and n_2 it turns out that an increase by 0.12 is sufficient for significance.

The figures and the examples illustrate the caveat of component comparisons based on recall, precision, and accuracy. For small sizes of n_1 and n_2 there must be a quite strong difference in the estimates to prove a difference in the components.

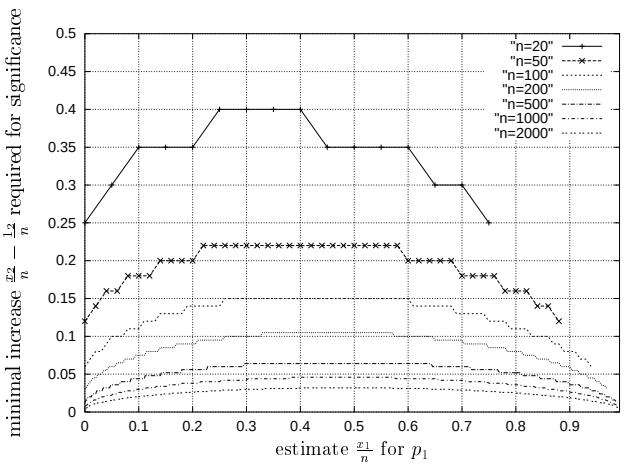

Figure 2. Minimal increase in maximum likelihood estimates needed for significance (error probability $\gamma = 5$%)

4 Conclusion

In this paper we have demonstrated the statistical background of evaluating document analysis components on test sets. Relying on the fact that recall, precision, and accuracy are maximum likelihood estimates of unknown probabilities, we have analysed two central goals pursued in experimental evaluations:

- to provide a guaranteed lower bound for the real effectiveness of a component and
- to compare two components based on a test set.

Certainly the statistical analysis presented here is not a contribution from the mathematical perspective. In contrast, the type of analysis presented here originates from the design of medical and psychological experiments and is much older than document analysis itself [4, 2, 5]. Nevertheless, researchers in document analysis often only have a rough impression of how their small test sets influence the experimental evaluation. Our diagramms illustrate how indicative experimental results in document analysis really are.

References

[1] H. Bunke and P. S. P. Wang, editors. *Handbook of Character Recognition and Document Image Analysis*. World Scientific Publishing Company, Singapore, 1997.

[2] C. Clopper and E. Pearson. Biometrika 26, 1934.

[3] A. Dengel and K. Hinkelmann. The Specialist Board — A Technology Workbench for Document Analysis and Understanding. In *Proceedings of the 2nd World Conference on Integrated Design and Process Technology (IDPT '96)*, pages 36–47, Austin, TX, USA, December 1996.

[4] R. Fisher. Inverse Probability. *Proceedings of the Campridge Philosphical Society*, XXVI(Pt. 4):528–535, 1930.

[5] R. Fisher. The Logic of Inductive Inference. *Journal of the Royal Statistical Society*, XCVIII(Pt. I):39–54, 1935.

[6] M. Junker and R. Hoch. An Experimental Evaluation of OCR Text Representations for Learning Document Classifiers. *International Journal on Document Analysis and Recognition*, 1(2):116–122, June 1998.

[7] W. Ledermann. *Handbook of Applicable Mathematics*, volume 6: Statistics. John Wiley & Sons, 1984.

[8] D. Lewis. *Representation and Learning in Information Retrieval*. PhD thesis, Department of Computer Science, University of Massachusetts, 1992.

[9] D. Michie, D. Spiegelhalter, and C. Taylor. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, 1994.

[10] G. Salton. *Automatic Text Processing; the Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Publishing Company, 1989.

[11] Y. Yang and J. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *Machine Learning. Proceedings of the 14th International Conference (ICML 97)*, pages 412–420, Nashville, TE, USA, July 6-12 1997.