Evaluation of Text Categorization Systems

October 3, 2011

Text categorization is one of the main and one of the youngest problems of information retrieval, however considering the arborescent research being started in the last 14 years, we can safely state that also the most explored one. In order to be able to test the performance of text categorization systems and more importantly to compare different systems against each other researchers applied and developed some measures for this task. The purpose of the present paper is to describe the classical measures applied for evaluating the performance of text categorization systems.

1 Text categorization

Categorizing textual data is the task of determining the correct class or classes of unseen documents based on some learning examples. Categories or classes are predetermined therefore a supervised learning algorithm is needed to classify documents.

Formally, following the notations used by Sebastiani in [3], let $\mathcal{D} = \{d_1, \ldots, d_{|\mathcal{D}|}\}$ and $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ be the set of the documents and categories respectively. The task is to approximate as precisely as it is possible the unknown target function by $\widetilde{\Phi} : \mathcal{D} \times \mathcal{C} \to \{T, F\}$, where T and F denotes "True" and "False", respectively. If there are only two classes, then the categorization is called binary, while if $|\mathcal{C}| > 2$ is called multiclass. Moreover, if the categories are disjunctive or non-overlapping, that is a document belongs to exactly one category, is called single-label, while if the categories are overlapping, that is a document can belong to arbitrary number of categories, it is said multilabel.

A text categorization system can use hard or ranking policy. Hard categorization means that the classifier takes a hard decision about a document-category pair, while ranking categorization ranks the categories for a given document based on some distance metric. This metric is called *catego-*

rization status value, $CSV_i : \mathcal{D} \to [0,1]$, $i = 1, \ldots, |\mathcal{C}|$. The greater the CSV-value of a category for a document the better the document belongs to the category.

2 Precision and recall

Precision and recall are the most common measures for evaluating an information retrieval system. Precision is the proportion of returned documents that are targets, while recall is the proportion of target documents returned.

Category i		Expert judgement	
		TRUE	**FALSE**
Classifier judgement	**TRUE**	TP_i	FP_i
	FALSE	FN_i	TN_i

Formally

$$P_i = \frac{TP_i}{TP_i + FP_i}$$

$$R_i = \frac{TP_i}{TP_i + FN_i}$$

The definition can be rewritten using probabilities. Suppose the learning algorithm was run and decisions had been made. This set constitutes our space. The precision can be written as

$$P_i = P(y = 1 \,|\, \widehat{y} = 1)$$

while recall is

$$R_i = P(\widehat{y} = 1 \,|\, y = 1)$$

Besides these accuracy and error another two common IR measures are sometimes used to characterize TC systems:

$$Acc_i = \frac{TP_i + TN_i}{TP_i + FP_i + FN_i + TN_i},$$

$$Err_i = \frac{FP_i + FN_i}{TP_i + FP_i + FN_i + TN_i}$$

There are two conventional methods of calculating the performance of a text categorization system based on precision and recall. The first is called *micro-averaging*, while the second one *macro-averaging*. Micro-averaged values are calculated by constructing a global contingency table (see above for

one class) and then calculating precision and recall using these sums. In contrast macro-averaged scores are calculated by first calculating precision and recall for each category and then taking the average of these. The notable difference between these to calculations is that micro-averaging gives equal weight to every document (it is called a *document-pivoted measure*) while macro-averaging gives equal weight to every category (*category-pivoted measure*).

$$P_{\text{micro}} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FP_i}; \quad R_{\text{micro}} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FN_i}$$

$$P_{\text{macro}} = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FP_i}; \quad R_{\text{macro}} = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FN_i}$$

3 Breakeven point

The properties or the expected behaviours of text categorization/information retrieval systems can vary. For example for one system it is better to return mostly correct answers, while in another it is better to cover more true positives. There is a tradeoff between precision and recall: if a classifier says "True" to every category for every document, then it receives perfect recall, but very low precision. However it can be easily seen that if a classifier says "False" for every category, except one which is correct ($TP = 1,\ FP = 0$) then it will have a precision equal to 1 but a very low recall. That is why it makes comparison between systems easier if the system is characterized by a single value, the *breakeven point* (BEP), which is the point at which precision equals recall. This can be achieved by tuning the parameters of the system. When there is no such point (because TP, FP and FN are natural numbers) the average of the nearest precision and recall is used, and is called *interpolated BEP*.

For example in ranking categorization models for each class an optimal τ_i CSV threshold has to be determined such that $P \simeq R$. If $CSV_i(d_j) \geq \tau_i$ then the classifier says "True", otherwise says "False".

4 11-point average precision

The 11-point average precision is another measure for representing performance with a single value. For every category the τ_i CSV threshold is repeatedly tuned such that allow the recall to take the values $0.0, 0.1, \ldots, 0.9, 1.0$.

At every point the precision is calculated and at the end the average over these eleven values is returned ([3]). The retrieval system must support ranking policy.

Yang ([4]) gives the following detailed algorithm for the calculation of this value. The precision and recall values for a given document and a threshold is calculated as

$$P = \frac{\text{categories found and correct}}{\text{total categories found}}; \quad R = \frac{\text{categories found and correct}}{\text{total categories correct}}$$

1. For each document calculate the precision and recall at each position in the ranked list where a correct category is found.

2. For each interval between thresholds $0.0, 1.0, \ldots, 0.9, 1.0$ use the highest precision value in that interval as the "representative" precision value at the left boundary of this interval.

3. For the recall threshold of 1.0 the "representative" precision is either the exact precision value if such point exists, or the precision value at the closest point in terms of recall. If the interval is empty we use the default precision value of 0.

4. Interpolation: At each of the above recall thresholds replace the "representative" precision using the highest score among the "representative" precision values at this threshold and the higher thresholds.

5. Per-interval averaging: Average per-document data points over all the test documents at each of the above recall thresholds respectively. This step results in 11 per-interval precision scores.

6. Global averaging: Average of the per-interval average precision scores to obtain a single-numbered performance average.

The resulting value is called the 11-point average precision.

5 The E and F-measures

Van Rijsbergen ([2], Ch.7) introduced two measures for evaluating information retrieval systems: the E and the F-measure. Since then the F-measure has become the most frequently used such measure.

The E-measure introduced by van Rijsbergen measures the non-overlapping grade between the retrieved and the actually relevant documents. Let **A** and **B** denote the set of relevant and the set of retrieved

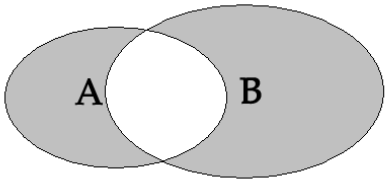

Figure 1: The graphical representation of the document sets. **A** denotes the relevant, while **B** the set of retrieved documents

documents, respectively (see Fig. 1). Lets denote the shaded region by $\mathbf{A}\Delta\mathbf{B} = \mathbf{A} \cup \mathbf{B} - \mathbf{A} \cap \mathbf{B}$. Then, by definition,

$$E = \frac{|\mathbf{A}\Delta\mathbf{B}|}{|\mathbf{A}| + |\mathbf{B}|},$$

where the denominator stands for normalizing the value, because we are interested only in the proportion of the relevant and non-relevant documents. Moreover from the contingency table in Section 2 we can write $|\mathbf{A}| = TP + FN$ and $|\mathbf{B}| = TP + FP$. Using these the E-measure becomes

$$E = \frac{FN + TP}{2TP + FP + FN} = 1 - \frac{2TP}{2TP + FP + FN} = 1 - 2 \cdot \frac{1}{\frac{TP+FP}{TP} + \frac{TP+FN}{TP}}$$

$$= 1 - \frac{1}{\frac{1}{2}\frac{1}{P} + \frac{1}{2}\frac{1}{R}} = \frac{2PR}{P+R},$$

which is the harmonic mean of the precision and recall. Thus, the more efficient the system is, the lower the E-value we obtain. The measure achieves its minimum at the highest values such that $P = R$ or $P \simeq R$ if such a point does not exist.

The F-measure is simply the complement of the E-measure, indicating the extent of overlap of the above sets, that is

$$F = \frac{2PR}{P+R}$$

A more generalized version of these are the E_β and F_β-measures, introducing a parameter $\beta \in [0, \infty)$ as a weighting factor for the importance of the recall (or precision) ([2], [1]):

$$E_\beta = 1 - \frac{(\beta^2+1)PR}{\beta^2 P + R}; \quad F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

By setting β to 1, we obtain the above E and F measures, that is why the conventional notation uses E_1 and F_1 denoting them. The most frequently applied setting is using $\beta = 1.0$. If $\beta = 0.5$, it means that recall is half as important as precision, in case of $\beta = 1.0$ they get equal importance, while for example $\beta = 2.0$ is used to set recall twice as important as the precision.

6 The relation of the BEP and the F-value

Because the BEP and the F-value captures the tradeoff between precision and recall, and using these a single value can be used for characterizing system performance, the researchers by choice use one these. But since there is no convention about which one to use, different papers use different measures. Both measures achieve its maximum at the highest values such that $P \simeq R$, in which case the BEP and the F-value are also equal. This section analyses the relation between them, which will be denoted by a bold question mark, **?**.

$$\frac{P+R}{2} \ \mathbf{?} \ \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

By multiplying the corresponding parts we get

$$\beta^2 P^2 - (\beta^2+1)PR + R^2 \ \mathbf{?} \ 0$$

Now if $P = 0$, then also $R = 0$ (from the definition), therefore **?**= " = ". If $R = 0$, then also $P = 0$, hence similarly **?**= " = ". In the case $P \neq 0$ and $R \neq 0$, we can proceed in the following way: divide both sides of the relation by R^2, and denote $\frac{P}{R} =: T$,

$$\beta^2 T^2 - (\beta^2+1)T + 1 \ \mathbf{?} \ 0$$

This can be written in the form of a product, that is

$$(T-1) \cdot (\beta^2 T - 1) \ \mathbf{?} \ 0$$

Now we can differentiate three cases:

1. $\boxed{P = R} \Rightarrow T = 1$, thus **?**= " = "
2. $\boxed{P < R} \Rightarrow T < 1$, thus $T - 1 < 0$ and $\beta^2 T - 1 < 0$, therefore **?**= " > "
3. $\boxed{P > R} \Rightarrow T > 1$, hence $T - 1 > 0$. Deciding the sign of the second component is somewhat longer. Solving the equation $f(\beta) = \beta^2 T - 1 = 0$ we get $\beta = \pm\frac{1}{\sqrt{T}}$. The function achieves its minimum at $\beta = 0$, so in $(-\infty, 0)$ monotonically decreases, and in $(0, \infty)$ monotonically increases, and $f(0) = -1$. Therefore

- if $\beta \in (-\frac{1}{\sqrt{T}}, \frac{1}{\sqrt{T}}) \Rightarrow \beta^2 T - 1 < 0$, that is **?**= " < "
- if $\beta \notin (-\frac{1}{\sqrt{T}}, \frac{1}{\sqrt{T}}) \Rightarrow \beta^2 T - 1 \geq 0$, i.e. **?**= " $\geq$ ".

So, unless $\beta \in (-\frac{1}{\sqrt{T}}, \frac{1}{\sqrt{T}})$, the BEP is greater or equal to F_β.

Because F_1 is just a specific setting of F_β this is also valid for it. But if we examine the third case (because the other cases do not depend on β), we see that $\beta \notin (-\frac{1}{\sqrt{T}}, \frac{1}{\sqrt{T}})$, because $\beta = 1 > \frac{1}{\sqrt{T}}$, which means that

$$\frac{P+R}{2} \geq \frac{2PR}{P+R},$$

that is BEP $\geq F_1$ (this could be shown by simply recalling that the arithmetic mean is always greater or equal than the harmonic mean). From these it follows that if the performance of an information retrieval system is computed using F_1, it will show a lowest performance unless $P = R$.

References

[1] David D. Lewis. Evaluating and optimizing autonomous text classification systems. In *SIGIR '95*, pages 246–254, New York, NY, USA, 1995. ACM Press.

[2] Cornelis J. Van Rijsbergen. *Information Retrieval*. Butterworths, 1979.

[3] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[4] Yiming Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1-2):69–90, 1999.