Feature Selection and Negative Evidence in Automated Text Categorization

[Poster Paper] *

Luigi Galavotti
AUTON S.R.L.
Via Jacopo Nardi, 2
50132 Firenze, Italy
galavott@tin.it

Fabrizio Sebastiani
Istituto di Elaborazione dell'Informazione, C.N.R.
56100 Pisa, Italy
fabrizio@iei.pi.cnr.it

Maria Simi
Dipartimento di Informatica, Università di Pisa 56125 Pisa, Italy
simi@di.unipi.it

ABSTRACT
We tackle two different problems of *text categorization*, namely feature selection (FS) and classifier induction. We propose a new FS technique, based on a simplified version of the χ^2 statistics and a novel variant, based on the exploitation of negative evidence, of the well-known k-NN method. We report the results of systematic experimentation of these two methods performed on the REUTERS-21578 benchmark.

1. INTRODUCTION
Text categorization denotes the activity of automatically building, by means of machine learning techniques, automatic text classifiers (see e.g. [2]). Two key steps are document indexing and classifier induction.

Document indexing refers to the task of automatically constructing internal representations of the documents, able to synthetize the meaning of the documents. Usually, a text document is represented as a vector of weights $d_j = \langle w_{1j}, \ldots, w_{rj} \rangle$, where r is the number of features (i.e. words) that occur at least once in at least one document of the collection, and $0 \leq w_{kj} \leq 1$ represents, loosely speaking, how much feature t_k contributes to the semantics of document d_j. Since classifier induction methods have a high computational cost, which is a function of r, *feature selection* techniques are used to select, from the original set of r features a subset of $r' \ll r$ features that are most useful for representing the meaning of the documents.

Classifier induction refers to the inductive construction of a text classifier from a training set of documents that have already undergone indexing and FS.

*A full version of this paper is available in *Proc. of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries, Lisbon, PT, 2000.*

2. FEATURE SELECTION
In feature selection the issue is to obtain a substantial reduction of the features, without compromising the performance of the classifier. The value $(1-\frac{r'}{r})$ measures the *aggressivity* of the selection: the higher this value, the smaller the set resulting from FS, and the higher the computational benefits.

A widely used approach to FS is the so-called *filtering* approach, which consists in selecting the r' features that score highest according to a function that measures the "importance" of the feature for the categorization task. Comparative experiments, performed across different induction methods and different benchmarks [5], have shown

$$\chi^2(t_k, c_i) = \frac{g[P(t_k, c_i)P(\overline{t}_k, \overline{c_i}) - P(t_k, \overline{c_i})P(\overline{t}_k, c_i)]^2}{P(t_k)P(\overline{t}_k)P(c_i)P(\overline{c_i})} \quad (1)$$

to be one of the most effective functions for the filtering method, allowing aggressivity levels in the range [.90,.99] with no loss (or even with a small increase) of effectiveness.

In Equation 1, g indicates the cardinality of the training set, and probabilities are interpreted on an event space of documents (e.g. $P(\overline{t}_k, c_i)$ estimates the probability that, for a random document x, feature t_k does not occur in x and x belongs to category c_i).

Intuitively, low values of $\chi^2(t_k, c_i)$ denote a high degree of independence of a feature t_k with respect to category c_i; thus we select those features for which $\chi^2(t_k, c_i)$ is highest.

Ng et al. [1] propose to take the square root of this formula, in order to emphasize the role of positive correlation between t_k and c_i (i.e. $P(t_k, c_i)$ and $P(\overline{t_k}, \overline{c_i})$) and de-emphasize negative correlation.

Conforted by their experimental results, which show a superiority of their variant over $\chi^2(t_k, c_i)$, we go a further step in this direction, by observing that:

- The $\sqrt{g}$ factor at the numerator is redundant, since it is equal for all pairs (t_k, c_i).
- The presence of $\sqrt{P(t_k)P(\overline{t}_k)}$ at the denominator emphasizes very rare features, since for these features it

has very low values. By showing that document frequency is a very effective FS technique, [5] has shown that very rare features are the least effective in TC.

- The presence of $\sqrt{P(c_i)P(\overline{c}_i)}$ at the denominator emphasizes very rare categories, since for these categories this factor has very low values. This in turn tends to depress microaveraged effectiveness, which is now considered the correct way to measure effectiveness in most applications [2].

Removing these three factors from the square root of $\chi^2(t_k, c_i)$ yields

$$s\chi^2(t_k, c_i) = P(t_k, c_i)P(\overline{t}_k, \overline{c}_i) - P(t_k, \overline{c}_i)P(\overline{t}_k, c_i) \quad (2)$$

which will be the basis for our experiments.

3. USING NEGATIVE EVIDENCE IN K-NN

One of the most popular paradigms for the inductive construction of a classifier is the *instance-based* approach, which is well exemplified by the k-NN (for "k nearest neighbors") algorithm used e.g. by Yang [3]. For deciding whether d_j should be classified under c_i, k-NN selects the k training documents most similar to d_j. Those documents d'_z that belong to c_i are seen as carrying evidence towards the fact that d_j also belongs to c_i, and the amount of this evidence is proportional to the similarity between d'_z and d_j. Classifying a document with k-NN thus means computing

$$CSV_i(d_j) = \sum_{d'_z \in Tr_k(d_j)} RSV(d_j, d'_z) \cdot v_{iz} \quad (3)$$

- $CSV_i(d_j)$ (*Categorization Status Value*) measures the computed evidence that d_j belongs to c_i;
- $RSV(d_j, d'_z)$ (*Retrieval Status Value*) is a measure of semantic relatedness between d_j and d'_z;
- $Tr_k(d_j)$ is the set of the k training documents d'_z with the highest $RSV(d_j, d'_z)$;
- the value of v_{iz} is given by

$$v_{iz} = \begin{cases} 1 & \text{if } d'_z \text{ is a positive instance of } c_i \\ 0 & \text{if } d'_z \text{ is a negative instance of } c_i \end{cases}$$

The threshold k is usually determined experimentally on a validation set; Yang [3, 4] has found $30 \leq k \leq 45$ to yield the best effectiveness.

Usually, the construction of a classifier also involves the determination of a threshold τ_i such that $CSV_i(d_j) \geq \tau_i$ may be viewed as an indication to file d_j under c_i. For determining this threshold we have used the *proportional thresholding* method, as in our experiments this has proven superior to *CSV thresholding* [2].

The basic philosophy that underlies k-NN and all the instance-based algorithms used in the TC literature is that only positive instances of a category are used as evidence towards the fact that d_j belongs to c_i.

We propose a variant of the k-NN approach in which evidence provided by negative training instances is also used. Mathematically, this comes down to using $v_{iz} = -1$ (instead of 0), if d'_z is a negative instance of c_i. We call the method deriving from this modification k-NN$_{neg}$.

We also tried variants of k-NN$_{neg}$, where the RSV component is raised to the power of 2 or 3, with the intention of limiting the negative contribution of very dissimilar training instances.

4. EXPERIMENTAL RESULTS

Experiments performed on the REUTERS-21578 benchmark have shown that simplified χ^2 has systematically outperformed χ^2, at very aggressive levels of reduction (from 0.960 to 0.999), and has done so by a wide margin. This fact, together with its low computational cost, make simplified χ^2 a very attractive method in those applications which demand radical reductions in the dimensionality of the feature space.

Concerning k-NN$_{neg}$, our hypothesis that evidence contributed by negative instances could provide an effectiveness boost for the TC task has been only partially confirmed by the experiments. In fact, our k-NN$_{neg}$ method has performed as well as the original k-NN but no better than it, and has furthermore shown to be more sensitive to the choice of k than the standard version. However, we have shown that with the variants of the original formula, which appropriately de-emphasize the importance of very dissimilar training instances, this method consistently outperforms standard k-NN. Given the prominent role played by k-NN in the TC literature, and given the simple modification that moving from k-NN to k-NN$_{neg}$ requires, we think this is an interesting result.

5. REFERENCES

[1] H. T. Ng, W. B. Goh, and K. L. Low. Feature selection, perceptron learning, and a usability case study for text categorization. In N. J. Belkin, A. D. Narasimhalu, and P. Willett, editors, *Proceedings of SIGIR-97*, 67–73, Philadelphia, US, 1997. ACM Press, New York, US.

[2] F. Sebastiani. Machine learning in automated text categorisation: a survey. Technical Report IEI-B4-31-1999, Istituto di Elaborazione dell'Informazione, C.N.R., Pisa, IT, 1999. .

[3] Y. Yang. Expert network: effective and efficient learning from human decisions in text categorisation and retrieval. In W. B. Croft and C. J. van Rijsbergen, editors, *Proceedings of SIGIR-94*, 13–22, Dublin, IE, 1994. Springer Verlag, Heidelberg, DE.

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

[5] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In D. H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.