ON STATISTICAL CLASSIFICATION OF SCIENTIFIC TEXTS

R. Rudzkis, V.Balys
Institute of Mathematics and Informatics
Vilnius, Lithuania
e-mail: rudzkis@ktl.mii.lt, vbalys@gmail.com

Abstract

The research considers the problem of classification of scientific texts. Models and methods based on stochastic distribution of scientific terms are discussed. The preliminar results of experimental study over real-world data are reported.

1 Introduction

The researches in the fields of scientific text processing, knowledge discovery and management face increasing attention in the light of ongoing changes of information finding and acquisition models. A piece of information or knowledge is of any value only if it can be easily and convienently found and reused. Therefore, the methods, algorithms and tools to perform tasks of indexing and classifying texts, organizing interactive search for content, extracting and delivering knowledge, and many other ones, are outright crucial both for publishers and the scientific society. Current research deals with such problems while in this thesis the main focus is put on classification of scientific texts which is one of the main problems in the field.

2 The Model

Definitions. Here, a rather brief introduction of stochastic term distribution models (that mathematically define the concept of identification cloud, [4]) based approach to the problem of classification is presented. For more comprehensive and detailed review see [5].

Let us consider the problem of classifying the scientific text (often it is only abstract), i.e., assigning one or maybe a few predefined labels to it. These labels may be classifiers from some classification system like MSC2000 or just plain keyphrases from controlled vocabulary. Actually, keyphrases should be treated differently as their vocabulary usually is very large and the procedure of assigning involves a substantial amount of randomness. However, this paper will not cover these details.

Let K denote some classification system of scientific texts which is identified with a set of all possible labels of the classes in that system. Let V be a vocabulary, i.e., set of scientific terms of a certain scientific field that are relevant to the classification of texts. The chronologically numerated vector of article's a elements $(a_1, \ldots, a_d)$, $d = d(a)$, where $a_i \in V$ and not necessarily $a_i \neq a_j$, is called the projection of the article a. Sometimes it is convenient to identify the projection of an article a with an

infinite sequence $(a_1, a_2, \ldots)$, where $a_i = 0$ for all $i > d(a)$. Here $0 \in V$ denotes an additional zero term which does not exist in reality. Let A be a set of projections of all articles or other publications from a certain scientific field. In what follows the word "projection" is omitted and $a \in A$ is called just an article.

From the point of view of classification an article is not necessarily a homogenous piece of text — in the general case, it consists of $q = q(a) \geq 1$ continuous homogenous parts which are classified as different in system K. Non-intersecting intervals of indices $I_j(a) \subset \{1, 2, \ldots, d(a)\}$ and class labels $w_j(a) \in K, \quad j = 1, \ldots, q$ correspond to these parts. Here $\bigcup_{j=1}^{q} I_j(a) = N(a)$ and $w_j \neq w_{j-1}, j = \overline{2, q}$: if two adjacent parts of the text are attributed to the same class they must be joined into one.

Let N be the set of natural numbers. An article $a \in A$ and a set of indices $I \subset N$ are chosen randomly so that the part of an article $\{(a_\tau, \tau), \tau \in I\}$ is homogenous: $I \subset I_\nu(a), \nu \in \{1, \ldots, q\}$. This part is attributed to the class $\eta = w_\nu(a)$ in the system K. A common problem of classification is to determine the unknown class η using the observed vector $a_I = (a_\tau, \tau \in I)$.

Probability distributions. Since (a, I, η) is the result of a random experiment, the probability distribution in the set K is defined by

$$Q(w) = \mathbb{P}\{\eta = w\}, \quad w \in K \ . \tag{1}$$

Let Y be a set of all possible values of a_I. In the set Y the following conditional probability distributions are defined:

$$P(y) = \mathbb{P}\{a_I = y \big| \ |I| = d(y)\} \ , \tag{2}$$

$$P(y|w) = \mathbb{P}\{a_I = y \big| \ |I| = d(y), \eta = w\}, \quad w \in K \ , \tag{3}$$

where $d(y) = dim\ y, |I| = card\ I$.

If η and $|I|$ are independent, after observing a_I, the a posteriori probability of the random event $\{\eta = w\}$ is determined by the equality

$$Q(w|a_I) = Q(w) \cdot \psi_w(a_I) \ , \tag{4}$$

where

$$\psi_w(y) = P(y|w)/P(y), y \in Y \ . \tag{5}$$

The concept of identification cloud may be defined by the functional ψ_w that reflects how the probability for the random text to belong to the class w depends on the frequency of terms in the text as well as on their positions.

Using the distributions, described in (1) and (3), Bayes classifier which minimizes mean classification losses can be defined. If the loss function is trivial, i.e., it equals to some constant in case of misclassification, it is simply the maximum aposteriori classifier:

$$\widehat{\eta} = \arg\max_{w \in K} P(a_I|w)Q(w) \ . \tag{6}$$

In equality (6) $\psi_{(\cdot)}(a_I)$ can be substituted for $P(a_I|\cdot)$.

Inference. In order to use this classification method, the distributions P and Q and the functional ψ, used in (6) must be estimated. Having the learning sample of observed parts of texts and their classification results $X = (y(1), \eta(1)), \ldots, (y(n), \eta(n))$, where $\eta(i) \in K, y(i) \in Y$, $Y = \{y = (y_1, \ldots, y_d) : y_i \in V, d \in N\}$, one can use non-parametric or parametric methods to get the estimates (see [5]).

3 Extending the model

The proposed model has a number of shortcomings. Some of them are related to the high dimensionality of feature space. The vocabulary of terms V is very large, even though it is much smaller than the vocabulary of all language words so there are too much parameters to estimate and having not so much learning data one would encounter overfitting problem which actually is the common problem in the field. There is a solution — the so called feature space dimensionality reduction technique where feature space V is mapped into other space U of much lower dimensionality where elements $u \in U$ represent some latent (unobserved) abstract concepts of higher level than single words or phrases $v \in V$. There is a number of different approaches ranging from rather heuristic method like Latent Semantic Indexing ([3]) to generative modeling of corpus like Latent Dirichlet Analysis ([1]). The applicability of these methods in the context of our models must be experimentally evaluated.

4 Experimental evaluation

There is a number of other plain text classification algorithms that gained high popularity over years, including naive Bayes, Support Vector Machines, k Nearest Neighbours and others (see [6] for a nice introduction into them). These algorithms may well be applied for classifying scientific texts, therefore, the results of the comparative performance study will be reported.

The experiments are being conducted on basis of more than 14000 articles from the field of probability theory and mathematical statistics kindly provided by the Institute of Mathematical Statistics, USA. The analysis is not finished yet and the final conclusions are still to be drawn. Below, there are some preliminar results and comments (mostly of qualitative nature due to the space restrictions) presented.

44 keyphrases and 92 MSC2000 classifiers were chosen for the experiments each having learning set of at least 50 abstracts, thus resulting in more than 2000 abstracts for keyphrases and more than 6000 abstracts for MSC2000 classifiers. Dictionary of near to 5000 most common scientific terms was used. 10-fold cross-validation was implemented to evaluate the performance of algorithms. Some of the analysed algorithms are: naive Bayes with Laplace smoothing, k Nearest Neighbours with cosine distance function and 20 neighbours (chosen as optimal by cross-validation), Support Vector Machines (with somewhat default parameters of LIBSVM [2]), Identification Clouds algorithm with conditional independence and stationarity assumption and rather small clouds (approximately 15 terms for an item), etc. Latent Semantic Indexing ([3]) was

implemented for feature space dimensionality reduction with various numbers of latent 'factors', e.g., 50, 100, 200, etc.

The simulation results of straightforward, i.e., without dimensionality reduction, algorithms differ unsignificantly with the Idenfication Clouds and SVM performing the best by a narrow margin (5 - 10 %). All the algorithms performed rather poorly reaching modest 55 - 60 % recall for classification strategy that chooses 2 highest ranked keyphrases for an article and up to 70 % for strategy that chooses 3 highest ranked keyphrases (on average article has 1.4 keyphrase assigned by authors). Using LSI seems to improve results but it is still to be estimated by how much.

More experiments are on the way. They will cover a wide selection of algorithm parameters as well as a number of additional methods. The results will be reported in the conference and presented in following papers.

References

[1] Blei D., Ng A., Jordan M. (2003). Latent Dirichlet allocation. *Journal of machine Learning Research*. Vol. **3**, pp. 993-1022.

[2] Chang C., Lin C. (2001). LIBSVM: a library for support vector machines. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[3] Deerwester S.C., Dumais S.T., Landauer T.K., Furnas G.W., Harshman R.A. (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*. Vol. **41**, pp. 391-407.

[4] Hazewinkel M. (2004). Dynamic stochastic models for indexes and thesauri, identification clouds, and information retrieval and storage. In: *R.Baeza-Yates a.o. (ed), Recent Advances in Applied Probability*. pp. 181-204.

[5] Rudzkis R., Balys V., Hazewinkel M. (2006). Stochastic Modelling of Scientific Terms Distribution in Publications. *Proceedings of 5th International Conference on Mathematical Knowledge Management, MKM2006, Lecture Notes in Artificial Intelligence*. Vol. **4108**, pp. 152-164.

[6] Sebastiani F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*. Vol. **34**, pp. 1-47.