MEDICAL DIAGNOSIS ASSISTANT BASED ON CATEGORY RANKING

Fernando Ruiz-Rico, Jose Luis Vicedo and María-Consuelo Rubio-Sánchez
University of Alicante, Spain

Keywords: Medical diagnosis, category ranking.

Abstract: This paper presents a real-world application for assisting medical diagnosis which relies on the exclusive use of machine learning techniques. We have automatically processed an extensive biomedical literature to train a categorization algorithm in order to provide it with the capability of matching symptoms to MeSH diseases descriptors. To interact with the classifier, we have developed a web interface so that professionals in medicine can easily get some help in their diagnostical decisions. We also demonstrate the effectiveness of this approach with a test set containing several hundreds of real clinical histories. A full operative version can be accessed on-line through the following site: www.dlsi.ua.es/omda/index.php.

1 INTRODUCTION

Text categorization consists of automatically assigning documents to pre-defined classes. It has been extensively applied to many fields such as search engines, spam filtering, etc. and in particular, some efforts have been focused on MEDLINE abstracts classification (Ibushi et al., 1999). However, as far as we are concerned, it has never been used to assist medical diagnosing by using the textual information provided by biomedical literature together with patient histories.

Every year, thousands of documents are added to the *National Library of Medicine* and the *National Institutes of Health* databases[1]. Most of them have been manually indexed by assigning each document to one or several entries in a controlled vocabulary called MeSH[2] (Medical Subject Headings). The MeSH tree is a hierarchical structure of medical terms which are used to define the main subjects that a medical article or report is about. Due to the wide use of this terminology, we can find translations into several languages such as Portuguese and Spanish (i.e. DeCS[3] - Health Science Descriptors). The diseases sub-tree not only defines on its own more than 4,000 pathological states, but also offers the chance to search for documented case reports related to each of them.

[1]http://www.pubmed.gov

[2]http://www.nlm.nih.gov/mesh

[3]http://decs.bvs.br/I/homepagei.htm

Our proposal tries to estimate a ranked list of diagnoses from a patient history. To tackle this problem, we have selected an existing categorization algorithm, and we have trained it using the textual information provided by lots of previously reported cases. This way, a detailed symptomatic description is sufficient to obtain a list of possible diseases, along with an estimation of probabilities.

We have not used binary decisions from binary categorization methods, since they might leave some interesting MeSH entries out, which should probably be taken into consideration. Instead, we have chosen a category ranking algorithm to obtain an ordered list of all possible diagnoses so that the user can finally decide which of them better suits the clinical history.

In this paper, first of all, we will explain the way we have developed our experiments, including a full description of the sources and methods used to get both training and test data. Secondly, we will provide an example of a patient history and both the expected and provided diagnoses. We will finish by showing and commenting several evaluation results on.

2 PROCEDURES

We have extracted the training data from the *PubMed* database[1] by selecting every case reports on diseases written in English including abstract and related to humans beings. These documents were retrieved by using the "diseases category[MAJR]" query, where

Ruiz-Rico F., Luis Vicedo J. and Rubio-Sánchez M. (2008).
MEDICAL DIAGNOSIS ASSISTANT BASED ON CATEGORY RANKING.
In *Proceedings of the First International Conference on Health Informatics*, pages 239-241

[MAJR] stands for “MeSH Major Topic”, asking the system for retrieving only documents whose subject is mainly a disease. The query provided us with 483,726 documents[4] that we downloaded by sending them to a file in MEDLINE format. We automatically processed that file to obtain the titles and abstracts with their corresponding MeSH topics. This led us to 4,024 classes with at least one training sample each.

With respect to the test set, we have used 400 medical histories from the School of Medicine of the University of Pittsburgh (Department of Pathology[5]). Although, so far the web page contains 500 histories[4], not all of them are suitable for our purposes. There are some which do not provide a concrete diagnosis but only a discussion about the case, and some others do not have a direct matching to the MeSH tree. We downloaded the HTML cases and afterwards we converted them to text format by using from each document the title and all the clinical history, including radiological findings, gross and microscopic descriptions, etc. To get the expected output, we extracted the top level MeSH diseases categories corresponding to the diagnoses given on the titles of the “final diagnosis” files (dx.html).

To select a proper ranking algorithm, we have looked up the most suitable one through several decades of literature about text classification and category ranking. We have chosen the Sum of Weights (SOW) approach (Ruiz-Rico et al., 2006), that is more suitable than the rest for its simplicity, efficiency, accuracy and incremental training capacity. Since medical databases are frequently updated and they also grow continuously, we have preferred using a fast and unattended approach that lets us perform updates easily with no substantial performance degradation after incrementing the number of categories or training samples. The restrictive complexity of other classifiers such as SVM could derivate to an intractable problem, as stated by (Ruch, 2005).

To evaluate how worth our suggestion is, we have measured accuracy through three common ranking performance measures (Ruiz-Rico et al., 2006): Precision at recall = 0 ($P_{r=0}$), mean average precision (*AvgP*) and Precision/Recall break even point (*BEP*). Sometimes, only one diagnosis is valid for a particular patient. In these cases, $P_{r=0}$ let us quantify the mistaken answers, since it indicates the proportion of correct topics given at the top ranked position. To know about the quality of the full ranking list, we use the AvgP, since it goes down the arranged list averaging precision until all possible answers are covered. BEP is the value where precision equals recall, that

[4]Data obtained on February 14th 2007

[5]http://path.upmc.edu/cases

Order	Diseases	Probability	Expand
1	Nervous System Diseases	100	+
2	Neoplasms	93	+
3	Pathological Conditions, Signs and Symptoms	61	+
4	Congenital, Hereditary, and Neonatal Diseases and Abnormalities	25	+
5	Endocrine System Diseases	23	+
...	...	...	...

Figure 1: Example of the first level of a hierarchical diagnosis.

Nervous System Diseases

Nervous System Neoplasms

Central Nervous System Neoplasms

Order	Diseases	Probability	Expand
1	Brain Neoplasms	100	+
2	Central Nervous System Cysts	44	+
3	Meningeal Neoplasms	13	+
4	Spinal Cord Neoplasms	3	+

Order	Diseases	Probability
1	Craniopharyngioma	100
2	Cysts	63
3	Pituitary Neoplasms	59
4	Central Nervous System Cysts	48
5	Headache	46
6	Xanthomatosis	44
...	...	...

Figure 2: Output example after manual expansion of high ranked topics (up) and by selecting the flat diagnosis mode (down).

is, when we consider the maximum number of relevant topics as a threshold. To follow the same procedure as (Joachims, 1998), the performance evaluation has been computed over the top diseases level.

2.1 Availability and Requirements

No special hardware nor software is necessary to interact with the assistant. Just an Internet connection and a standard browser are enough to access on-line through the following site: *www.dlsi.ua.es/omda/index.php*.

By using a web interface and by presenting results in text format, we allow users to access from many types of portable devices (laptops, PDA’s, etc.). Moreover, they will always have available the latest version, with no need of installing specific applications nor software updates.

3 AN EXAMPLE

One of the 400 histories included in the test set looks as follows:

Case 177 – Headaches, Lethargy and a Sellar/Suprasellar Mass

A 16 year old female presented with two months of progressively worsening headaches, lethargy and visual disturbances. Her past medical history included developmental delay, shunted hydrocephalus, and tethered cord release ...

The final diagnosis expected for this clinical history is: “Rathke’s Cleft Cyst”, which is a synonym of the preferred term “Central Nervous System Cysts”. Translating this into one or several of the 23 top MeSH diseases categories we are lead to the following entries:

- Neoplasms
- Nervous System Diseases
- Congenital, Hereditary, and Neonatal Diseases and Abnormalities.

In hierarchical mode, our approach provides automatically a first categorization level with expanding possibilities as shown in figure 1. We provide navigation capabilities to allow the user to go down the tree by selecting different branches, depending on the given probabilities and his/her own criteria. Moreover, a flat diagnosis mode can be activated to directly obtain a ranked list of all possible diseases, as shown in figure 2.

After an individual evaluation of this case, we have obtained the following values: $P_{r=0} = 1$, $AvgP = 0.92$, and $BEP = 0.67$, since the right topics in figure 1 are given at positions 1, 2 and 4.

4 RESULTS

Last row in table 1 shows the performance measures calculated for each medical history and its diagnosis, averaged afterwards across all the 400 decisions. $P_{r=0}$ indicates that we get 69% of the histories correctly diagnosed with the top ranked MeSH entry. AvgP value means that the rest of the list also contains quite valid topics, since it reaches a value of 73%.

First row in table 1 provides a comparison between SVM (Joachims, 1998) and sum of weights (Ruiz-Rico et al., 2006) algorithms using the well known OHSUMED evaluation benchmark. Even using a training and test set containing different document types, BEP indicates that the performance is not far away from that achieved in text classification tasks, meaning that category ranking can also be effectively applied to our scenario.

Table 1: Averaged performance for both text categorization and diagnosis.

Corpus	**Algor.**	$\mathbf{P_{r=0}}$	**AvgP**	**BEP**
OHSUMED	SVM	-	-	0.66
	SOW	-	-	0.71
Case reports and patient histories	SOW	0.69	0.73	0.62

5 CONCLUSIONS

We believe that category ranking algorithms may provide a useful tool to help in medical diagnoses from clinical histories. Although the output of the categorization process should not be directly taken to diagnose a disease without a previous review, the accuracy achieved could be good enough to assist human experts. Moreover, our implementation demonstrates that both training and classification processes are very fast, leading to an accessible and easy upgradable system.

REFERENCES

Ibushi, K., Collier, N., and Tsujii, J. (1999). Classification of medline abstracts. *Genome Informatics, volume 10*, pages 290–291.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142.

Ruch, P. (2005). Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics, volume 22 no. 6 2006*, pages 658–664.

Ruiz-Rico, F., Vicedo, J. L., and Rubio-Sánchez, M.-C. (2006). Newpar: an automatic feature selection and weighting schema for category ranking. In *Proceedings of DocEng-06, 6th ACM symposium on Document engineering*, pages 128–137.