Handling Negation in Classification of Clinical Texts

Jacinto Mata, Manuel J. Maña, José M. Bermúdez, Noa P. Cruz, Patricia Jiménez
Universidad de Huelva, Spain

Abstract

In this paper we describe our system focused on the management of the negation. Handling negation in narrative clinical documents is very important to classification systems since they must identify whether observations in a report are present or absent. In order to handle the negations in the data sets we have used NegEx program, based on regular expressions. In the other hand, it is easy to find several names to referring the same concept in medical texts. This characteristic requires the utilization of a controlled vocabulary that allows to identify linguistic variants or different phrases for a same concept. An evaluation on training data performing 3-fold cross-validation and repeating each run 5 times shows similar results to the mean results for the challenge. However, the evaluation on test data provided results that are significantly worse than the obtained in training phase. We think that it can be owing to an overfitting during the training process.

Introduction

Text classification is an important research problem in text mining area, which consists in assigning pre-defined classes to text documents[1]. Typically, the document space is some type oh high-dimensional space and the classes are human-defined for the needs of an application. Currently, there is an increasing interest in text mining and information extraction strategies applied to the biomedical and molecular biology literature due to the increasing number of electronically available publications.

Nowadays a great number of patient clinical records is stored in digital format as narrative reports (electronic health records). These reports describe observations, physical symptoms, and clinical assessments that can be useful for clinical research. An example of patient clinical records are the discharge summaries. Researchers in the medical informatics community have developed techniques for extracting information from the clinical data[2]. The information obtained form these narrative reports has been used for decision support, guideline implementation, detection and management of epidemics or identification of patients eligible for research studies.

The challenge proposed by the Informatics for Integrating Biology and the Bedside (i2b2) in its second edition was a multi-class, multi-label classification task focused on obesity and its co-morbidities. The goal was to evaluate systems on their ability to recognize whether a patient is obese and what co-morbidities they exhibit from narrative reports. In this case, the reports were discharge summaries.

In this paper we describe our system for the second i2b2 task. We directed most of our efforts to the management of the negation. Handling negation in narrative clinical documents is very important to classification systems since they must identify whether observations in a report are present or absent.

The remain of this paper is organized as follows. Section 2 describe the data sets provided by the organization. In Section 3 we present our system architecture and the resource we have used. The evaluation with training data and test data are shown in Section 4 and Section 5. Finally, we present some conclusions about our system.

Data sets

The data for the challenge consisted of discharge summaries from Partners Healthcare, previously de-identified. The aim was to discover whether a patient suffered some of the sixteen co-morbidities (including obesity).

The training set was annotated by two obesity experts from two ways: textual judgments were based strictly on assertions made in the text whereas intuitive judgments were based on the subjective judgment of the experts. They went through each record and for each co-morbidity assigned a label. For textual judgements the label were present (Y), absent (N), questionable (Q), or unmentioned (U) whereas for intuitive judgments only Y, N, and Q labels were annotated. An example of annotations for training records file for intuitive judgments is shown in Figure 1.

```
<diseaseset>
<diseases source="intuitive">
<disease name="Asthma">
<doc id="1" judgment="N"/>
<doc id="2" judgment="Y"/>
<doc id="4" judgment="N"/>
...
<doc id="1240" judgment="N"/>
<doc id="1242" judgment="N"/>
<doc id="1245" judgment="N"/>
<doc id="1248" judgment="Y"/>
<doc id="1249" judgment="Y"/>
</disease>
<disease name="CAD">
<doc id="2" judgment="N"/>
<doc id="4" judgment="Y"/>
<doc id="6" judgment="N"/>
<doc id="13" judgment="Y"/>
...
<doc id="1244" judgment="N"/>
<doc id="1245" judgment="Y"/>
<doc id="1246" judgment="Y"/>
<doc id="1249" judgment="N"/>
</disease>
```

Figure 1. Anotation XML file example.

The training set was made up of 730 discharge summaries. An important feature of the training set was the unbalanced distribution of the classes for both judgments and for all the co-morbidities. For example, for the co-morbidity "depression", the labels distribution was N (0 records), Q (0 records), U (624 records) and Y (104 records) for textual judgments and N (555 records), Q (0 records) and Y (142 records) for intuitive judgments.

Finally, the test set was made up of 507 discharge summaries with the same structure as the training set.

System architecture

The system architecture is shown in Figure 2. In an initial phase, a preprocessing task allows to extract, for each disease, its patient records and its class labels associated.

In medical texts, it is easy to find several names to referring the same concept. This characteristic requires the utilization of a controlled vocabulary that allows to identify linguistic variants or different phrases for a same concept. The MetaMap Transfer[3] software (MMTx) tokenizes documents into sentences, phrases, terms and words. Then, it matches the phrases to the closest UMLS Metathesaurus[4] concepts. The Metathesaurus is a very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them. When a phrase has multiple candidate concepts, all the most probable candidates have chosen.

The representation of a discharge summary as a bag-of-concepts is unable to express the semantic associated to the concept negations. For example, in the sentence "She denies tobacco use and drinks alcohol rarely" appears the concepts *tobacco* and *alcohol*. However, the presence of these concepts next to the negation *denies* shows that the patient does not smoke nor drink. A simple representation of this sentence as a bag-of-concepts would show the opposite meaning. Then it is necessary to identify automatically the negated concepts.

In order to handle the negations in the data sets we have used NegEx[5] program. NegEx is a simple algorithm based on regular expressions and a dictionary of medical concepts. We have used the Snomed dictionary available with NegEx. The algorithm handles two types of negations:

- *Type 1.* A negation phrase that occur before or after the clinical term which is being negated.
- *Type 2.* A phrase that occur before or after a clinical term and indicates conditional possibility.

Each patient record is represented as a row of a matrix where the columns are the concepts discovered by MMTx in the training data set and a column with the label class. We have used two sets of values to represent the patient records. In the first case, the matrix values can be {1, 0, -1, -2} with the following meaning:

- 1: the concept appears more times affirmatively than in negative forms of type 1.
- 0: the concept does not appear into the record patient.
- -1: the concept appears more times in negative forms of type 1 than in affirmative way.
- -2: the concept appears at least on time negated with negations of type 2.

In the second case, the set of values is {1, 0, -1, -2, -3} with the following meaning:

- 1: the concept appears at least one time and always in positive form.
- 0: the concept does not appear into the record patient.
- -1: the concept appears at least one time and always negated with negations of type 1.

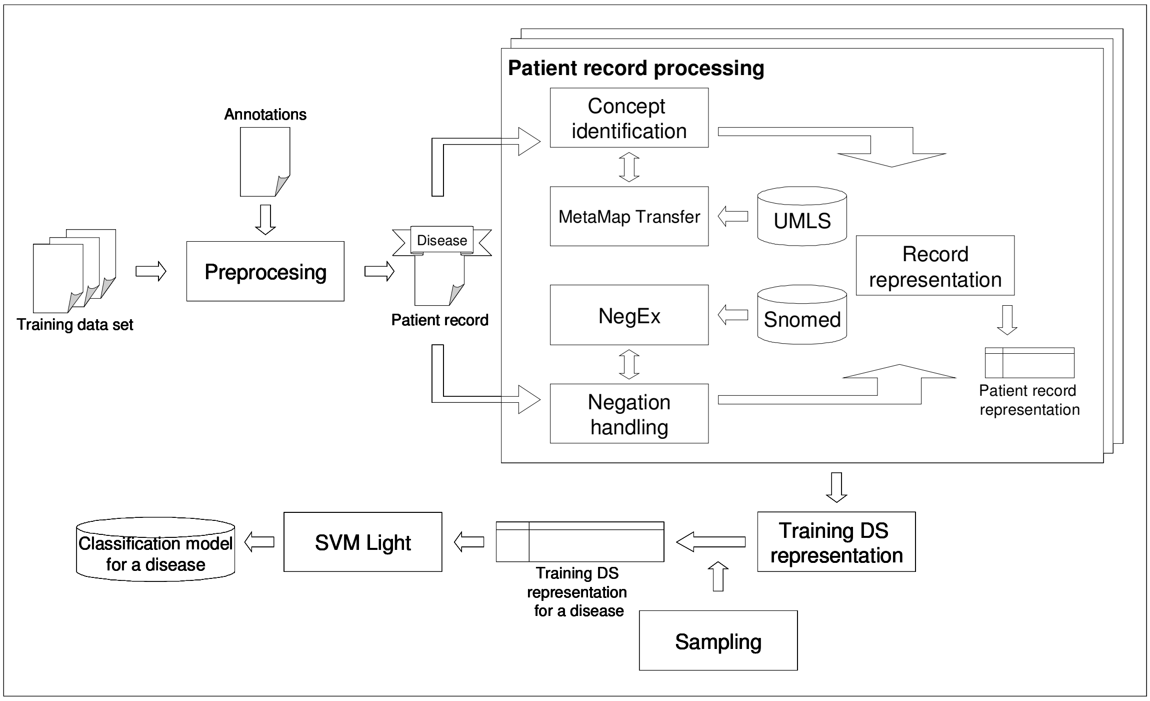

Figure 2. System architecture.

- -2: the concept appears at least on time negated with negations of type 2.
- -3: the concept appears into the record patient in both affirmative and negative ways (negations of type 1).

The representation of the training data set for a given disease is built from the representations of all the patient records which have a judgment for this disease.

In this problem, the training set has an unbalanced distribution of the classes, especially "N" and "Q" classes are poorly represented for both judgments and for all the co-morbidities. Table 1 shows the distribution of all classes in the training data.

Judgement	Q	N	Y	U
Textual	0.33	0.75	27.59	71.33
Intuitive	0.24	69.10	30.66	–

Table 1. Percentage of class distribution in training data set.

Data mining of unbalanced datasets will often involve adjustments to the modeling in some way[6]. Our approach is to over sample the minority cases and under sample the majority cases, using a random strategy to even up the classes. With this combination of strategies we obtained the distribution of classes shown in Table 2.

Judgement	Q	N	Y	U
Textual	10.0	20.0	30.0	40.0
Intuitive	10.0	50.0	40.0	–

Table 2. Percentage of class distribution after sampling.

Finally, we have used Support Vector Machines to build a classifier for each disease. More specifically, we have used the *SVM*^{light} implementation[7] for multi-class problems[8]. Then, we obtain 32 classifiers, i.e., 16 classifiers for each type of judgment.

Evaluation with training data

Using the scheme shown in Figure 2 we have built 42 different types of classification models. With the training data, we have built and evaluated these models performing stratified 3-fold cross-validation and repeating each run 5 times. Among these classifiers, we have selected those that achieve the best results.

The systems selected for textual judgment are:

- System 1. Identification of concepts using NegEx with Snomed. The set of values used to represent the patient records is {1, 0, -1, -2}.
- System 2. Identification of concepts using NegEx with Snomed. The set of values used to represent the patient records is {1, 0, -1,

-2}. We performed sampling of the training data.

- System 3. Identification of concepts using NegEx with Snomed. The set of values used to represent the patient records is {1, 0, -1, -2, -3}. We performed sampling of the training data.

The systems selected for intuitive judgment are:

- System 1. Identification of concepts using NegEx with Snomed. The set of values used to represent the patient records is {1, 0, -1, -2}.
- System 2. Identification of concepts using NegEx with Snomed. The set of values used to represent the patient records is {1, 0, -1, -2}. We performed sampling of the training data.
- System 3. Identification of concepts using NegEx with Snomed. The set of values used to represent the patient records is {1, 0, -1, -2, -3}.

Tables 3 and 4 show the averaged results of the evaluation using stratified cross-validation in the training data. For textual judgments, the system 2 achieves the best macro-averaged F-measure with a value of 0.6018. The best micro- averaged F-measure is achieved by system 1 with a value of 0.9143.

Run	Micro P	Macro P	Micro R	Macro R	Micro F	Macro F
1	0.9143	0.8779	0.9143	0.5897	**0.9143**	0.6016
2	0.8880	0.8634	0.8880	0.5930	0.8880	**0.6018**
3	0.8785	0.8497	0.8785	0.5851	0.8785	0.5955

Table 3. Performance for training data and textual judgments.

With regard to intuitive judgments, the system 1 achieves the best macro and micro-averaged F-measure with values of 0.7042 and 0.9041, respectively.

Run	Micro P	Macro P	Micro R	Macro R	Micro F	Macro F
1	0.9041	0.8495	0.9041	0.6913	**0.9041**	**0.7042**
2	0.8777	0.8362	0.8777	0.6953	0.8777	0.7033
3	0.8940	0.8344	0.8940	0.6804	0.8940	0.6936

Table 4. Performance for training data and intuitive judgments.

The results obtained by cross-validation of the training data are similar to the mean results for the challenge.

Evaluation with test data

Tables 5 and 6 show the performance obtained by the systems described above on the test data. As can be seen, the results are significantly worse than the obtained in training phase, in spite of have used cross-validation to build and evaluate the classifiers. We think that, perhaps, it can be owing to an overfitting during the training process.

Run	Micro P	Macro P	Micro R	Macro R	Micro F	Macro F
1	0.4233	0.2531	0.4233	0.2523	0.4233	0.2237
2	0.4247	0.2529	0.4247	0.2498	0.4247	0.2243
3	0.4310	0.2555	0.4310	0.2564	**0.4310**	**0.2275**

Table 5. Performance for test data and textual judgments.

Run	Micro P	Macro P	Micro R	Macro R	Micro F	Macro F
1	0.5274	0.3576	0.5274	0.3679	0.5274	**0.3442**
2	0.5257	0.3461	0.5257	0.3553	0.5257	0.3358
3	0.5289	0.3537	0.5289	0.3439	**0.5289**	0.3413

Table 6. Performance for test data and intuitive judgments.

Conclusion

In this paper, we show the outline of our systems. From this scheme, we have built 42 different types of classification models by using distinct strategies.

In medical reports a great number of negation phrases appear to be used the majority of the time to indicate the absence of clinical observations. Identifying whether the concepts are positive or negative is decisive to representing the information described in the report. Therefore, any system indexing clinical observations in narrative reports should handle negation.

Hence the best systems we have obtained have been those that have used the NegEx algorithm to identify the absence of clinical observations.

In the future we will continue working to improve the handling of the negation in clinical text. We are planning to use machine learning techniques instead of regular expression approach of NegEx.

References

1. Sebastiani F. Machine learning in automated text categorization. ACM Comput. Surv. 2002;34(1): 1–47.
2. Hripcsak G, Bakken S, Stetson PD, Patel VL. Mining complex clinical data for patient safety research: a framework for event discovery. J. of Biomedical Informatics. 2003;36(1/2):120–130.

3. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc. AMIA Symp. 2001;17–21.
4. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf. Med. 1993;32:281–291.
5. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan B. Evaluation of negation phrases in narrative clinical reports. Proc AMIA Symp. 2001;105–9.
6. Chawla NV, Japkowicz N, Kolcz A. Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations. 2004; 6(1): 1–6.
7. Joachims T. Learning to Classify Text Using Support Vector Machines. Dissertation, Kluwer, 2002.
8. Crammer K, Singer Y. On the Algorithmic Implementation of Multi-class SVMs. J. of Machine Learning Research. 2001;2(Dec):265–292.