Journal of Biomedical Informatics 87 (2018) 50–59

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Predicting of anaphylaxis in big data EMR by exploring machine learning approaches

Isabel Segura-Bedmar[a,*], Cristobal Colón-Ruíz[a], Miguél Ángel Tejedor-Alonso[b,c], Mar Moro-Moro[b]

[a] Computer Science Department, University Carlos III of Madrid, Avenida de la Universidad 30, 28911 Leganés, Madrid, Spain
[b] Allergy Unit, Hospital Universitario Fundación, Avenida Budapest 1, 28922 Alcorcón, Madrid, Spain
[c] Medicine and Surgery Department, Universidad Rey Juan Carlos, 28922 Alcorcón, Madrid, Spain

ARTICLE INFO

Keywords:
Machine learning
EMR classification
Bag of centroids
Anaphylaxis
Balancing strategies

ABSTRACT

Anaphylaxis is a life-threatening allergic reaction that occurs suddenly after contact with an allergen. Epidemiological studies about anaphylaxis are very important in planning and evaluating new strategies that prevent this reaction, but also in providing a guide to the treatment of patients who have just suffered an anaphylactic reaction. Electronic Medical Records (EMR) are one of the most effective and richest sources for the epidemiology of anaphylaxis, because they provide a low-cost way of accessing rich longitudinal data on large populations. However, a negative aspect is that researchers have to manually review a huge amount of information, which is a very costly and highly time consuming task. Therefore, our goal is to explore different machine learning techniques to process Big Data EMR, lessening the needed efforts for performing epidemiological studies about anaphylaxis. In particular, we aim to study the incidence of anaphylaxis by the automatic classification of EMR. To do this, we employ the most widely used and efficient classifiers in text classification and compare different document representations, which range from well-known methods such as Bag Of Words (BoW) to more recent ones based on word embedding models, such as a simple average of word embeddings or a bag of centroids of word embeddings. Because the identification of anaphylaxis cases in EMR is a class-imbalanced problem (less than 1% describe anaphylaxis cases), we employ a novel undersampling technique based on clustering to balance our dataset. In addition to classical machine learning algorithms, we also use a Convolutional Neural Network (CNN) to classify our dataset. In general, experiments show that the most classifiers and representations are effective (F1 above 90%). Logistic Regression, Linear SVM, Multilayer Perceptron and Random Forest achieve an F1 around 95%, however linear methods have considerably lower training times. CNN provides slightly better performance (F1 = 95.6%).

1. Introduction

Anaphylaxis is a life-threatening allergic reaction that occurs suddenly after contact with an allergen [1]. Despite its possible severe symptoms, even today there is no agreement about clinical criteria for diagnosing anaphylaxis. This lack of specific criteria can result in a failure to properly treat anaphylaxis, with fatal consequences for the patient, because of its early onset and very rapid progression. This has also translated into increased research into the epidemiology of this disorder [1], because epidemiological studies about anaphylaxis are very important in planning and evaluating new strategies that prevent this reaction, but also in providing a guide to the treatment of patients who have just suffered it.

Electronic Medical Records (EMR) are one of the most effective and richest sources for the epidemiology of anaphylaxis, because they provide a low-cost way of accessing rich longitudinal data on large populations. However, researchers have to manually review a huge amount of information, which is a very costly and highly time consuming task. Thus, our goal is to alleviate this burden to researchers, by providing them a system capable to automatically classify if a record describes a case of anaphylaxis or not. This automatic classification allows researchers to know, among other things, the incidence (total number of new cases) of anaphylaxis among a population.

Although the classification of EMR can be addressed by using a set of rules written by experts, this approach has several drawbacks. Firstly, its coverage is very low because of the richness of the human language

* Corresponding author.
E-mail address: isegura@inf.uc3m.es (I. Segura-Bedmar).

https://doi.org/10.1016/j.jbi.2018.09.012
Received 4 January 2018; Received in revised form 31 August 2018; Accepted 24 September 2018
Available online 25 September 2018
1532-0464/

with a wide variety of possible expressions to refer to the same object [2,3]. Moreover, when the number of rules is large, some rules can conflict with one another, hindering their maintenance. No less important is the fact that the rules cannot be reused when the classification problem changes. Unlike the rules-based systems, machine learning algorithms are domain-independent with high predictive performance [4]. In this work, we explore the most widely used and efficient machine learning classifiers in text classification to identify highly probable anaphylaxis cases in EMR. In addition to using the popular bag-of-words approach to represent the records, we use a novel method to represent them using a bag of centroids, which are calculated using the word embeddings from the records. One of the main advantages of our approach, which does not require any knowledge about anaphylaxis, is that it could easily be adapted to identify other diseases and disorders.

Several epidemiological studies put the incidence of anaphylaxis between 50 and 112 episodes per 100,000 person-years [5]. As results, the identification of anaphylaxis cases is a very class-imbalanced problem, which occurs when one class has many more training examples than the other class. This can negatively affect the performance of machine learning techniques, because they tend to be biased towards the majority class, but degrading their performance on minority class. Thus, it is necessary to handle this imbalance issue in order to improve the classification performance of the minority class. The most common approach to overcome this problem is to balance the data distribution on the dataset by using undersampling (decreasing the number of majority classes) and/or oversampling (cloning the minority class instances) methods. Several works [6,7] have proven that undersampling produces better results than oversampling, which tends to increase the chances of overfitting. On the other hand, representative instances of the majority class could be ignored by using undersampling.

An innovative solution to this drawback has been recently proposed in [8], where a clustering technique, in particular, the k-means clustering algorithm [9] is used to replace similar instances by their cluster centroid (this is explained with more detailed in Section 3). The authors validated their approach on several datasets from the areas of bioinformatics and quantum physics[1] and a dataset of X-ray images of the breast.[2] However, none of them was composed of texts. Therefore, this work is the first to validate the effectiveness of this undersampling method on an unbalanced dataset of texts.

To sum up, the main contributions of our work are as follows:

- According to the analysis of related work (see Section 2), we propose a bag-of-centroids approach to represent the records, which has never used in the clinical domain. We also explore and compare other ways to represent clinical records such as the BoW model, BoW with tf-idf and the average of word embeddings of all the words in a document.
- We apply a novel undersampling technique [8] based on the use of k-means clustering algorithm. This technique has never been applied before to balance text datasets.
- We provide a detailed analysis of different machine learning algorithms to classify EMR and find the optimum solution for the identification of anaphylaxis cases, with the final goal of reducing the heavy workload in epidemiological studies.
- Our approach does not require the use of domain knowledge, and thereby, can be easily extended to other text classification problems in the medical domain and languages other than Spanish (our dataset is composed of EMR written in Spanish).

The organization of this paper is as follows. In next section, we discuss previous works in EMR text classification. Section 3 presents the research methodology, including the description of the dataset and the representation of records as instances as well as the balancing technique and machine learning classifiers applied to identify anaphylaxis cases in EMR. In Section 4, we present and discuss the experimental results. Finally, conclusions and potential future work items are identified in Section 6.

2. Related work

The purpose of this section is to discuss the main works that benefit of using Natural Language Processing (NLP) in clinical domain, specifically, for classification of EMR. At the end of this section, we also briefly review the main data balancing techniques.

2.1. Classification of EMR using machine learning algorithms

In text classification, documents are represented by vectors of features, which are the input for machine learning classifiers. Many systems use the popular and simple BoW approach, where tokens (except most common words) are considered as features and are represented by their relative frequency. Other systems exploit NLP tools (such as PoS taggers, noun phrase chunkers or named entity taggers, among others) to obtain linguistic and semantic features to represent texts. Bellow we present a summary of recent work in EMR classification by using NLP and machine learning.

Machine learning and NLP methods were combined in [10] to identify EMR reporting a clinically important brain injury. The authors created an artificial dataset of 3621 EMR that describe normal and abnormal head computed tomography (CT) scans. The dataset was manually reviewed by doctors to identify those records that report a clinically important brain injury. Texts were tokenized and punctuation, stop-words and superfluous words were removed. Then, texts are represented with vectors of word frequencies using a BoW model. Three different classifiers were proposed due to its ability to determine the probability of its classifications: K-Nearest Neighbours algorithm (k-NN), Decision Tree classifier and Support Vector Machine (SVM), which obtained the best results (recall of 93.33% and precision of 50%).

EMR were used to classify patient alcohol use in [11]. The records were represented using a BoW approach, which captures the relative frequency of unigrams and bigrams in each document. The system used SVM classifier and was evaluated on a dataset of 2000 records manually classified, providing a very high performance for the detection of current drinkers (F1 = 89%).

The BoW model was also used in [12] to represent EMR (written in Georgian language). However, in this work, instead of having a binary classification problem, the records had to be classified in three different categories: ultrasonography, endoscopy, and X-ray. SVM and k-NN algorithms were applied, showing very close performance (F1 ranges from 82% to 88%, depending of the category).

Although the BoW model is very effective to represent texts, several works have used other features to represent EMR. Liu et al. [13] adapted the original cTAKES [14] smoking module by adding additional keywords (such as "anxiety" or "dependence") and some sentences (for example "raspy smoker's laugh", "smells of cigarette smoke"). Then, these keywords and sentences were used to train a SVM classifier with a radial basis kernel. The approach was trained and tested using a dataset of 400 clinical notes, providing a precision of 94% and a recall of 94%.

cTAKES was also used in [15] to process colonoscopy pathology reports in order to obtain their tokens, part of speech tags and negated terms. These features, along with a dictionary of medical and lay terms, were used to represent each report. A Conditional Random Forest (CRF) was trained to classify pathology reports as either derived from surveillance or non-surveillance colonoscopy in patients with inflammatory bowel disease (IBD). A dataset of 575 reports was manually classified by a gastroenterologist. The system had a precision of 80% and a recall of 77%.

[1] http://www.kdd.org/kdd-cup/view/kdd-cup-2004.
[2] http://www.kdd.org/kdd-cup/view/kdd-cup-2008.