ICD9-based Text Mining Approach to Children Epilepsy Classification

Luis Pereira\textsuperscript{a,1}, Rui Rijo\textsuperscript{a,b}, Catarina Silva\textsuperscript{a,c}, Margarida Agostinho\textsuperscript{d}

\textsuperscript{a}School of Technology and Management / Research Center for Informatics and Communications
Polytechnic Institute of Leiria - Portugal
\textsuperscript{b}INESCC - Institute for Systems and Computers Engineering at Coimbra - Portugal
\textsuperscript{c}CISUC – Centre for Informatics and Systems Of the University of Coimbra - Portugal
\textsuperscript{d}Centro Hospitalar de Leiria-Pombal, Hospital Santo André

Abstract

According with the World Health Organization, around 50 million people in the world have epilepsy. After the diagnosis process, physicians classify epilepsy according to the International Classification of Diseases, Ninth Revision (ICD-9). Often exams as electroencephalograms and magnetic resonances are used to create a more accurate diagnosis in a short amount of time. The classification process is time consuming and demands the realization of complementary exams. To circumvent this laborious process we propose an automatic process of classifying epileptic diagnoses based on ICD-9. We put forward a text mining approach, using processed electronic medical records and a K-Nearest Neighbor is applied as a white-box multi classifier approach to classify each instance mapping into the corresponding standard code. Results suggests a good performance proposing a diagnosis from electronic medical records, despite of the reduced volume of available training data.

© 2013 The Authors Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
Selection and/or peer-review under responsibility of SCIKA – Association for Promotion and Dissemination of Scientific Knowledge
Keywords: data mining, text mining, electronic medical records, ICD codes, machine learning, epilepsy;

\textsuperscript{1} Corresponding author. Tel.: +351910563111.
E-mail address:luispereira268@gmail.com.
1. Introduction

In Portugal, and according to Intercontinental Medical Statistics (IMS), epilepsy is the second most common neurological disease, affecting 70,000 people each year (from a 10 million population) and 50 million people around the world.

The process of identifying and classifying epilepsy is complex, demanding considerable time and effort [1]. Several characteristics, symptoms and exams need to be taken into account to reach a precise epilepsy diagnosis, to define the procedures and the medication according to the epileptic seizure type. In fact, some types of epilepsy need a rapid action in order to control the seizures and allow a normal and productive quotidian. Moreover, this difficulty becomes intensified in children, because it requires the analysis of different causes, e.g. genetic, structural or metabolic. The responsibility of such a diagnosis in children is overwhelming because it can dramatically change a child’s life, and a misdiagnose of epilepsy can lead to an inefficient therapy or even become fatal if not identified and controlled appropriately [2]. These diagnoses must be classified according a standard code as the International Classification of Diseases, Ninth Revision (ICD-9), currently in use in Portugal and in other countries like the United States.

ICD-9 codes are used to describe a patient’s diagnosis including symptoms, diseases or disorders. An ICD-9 code means that every medical professional in Portugal and many other parts of the world will understand the diagnosis the same way. It is important for ICD-9 codes to be accurate for the quality of patient care and to prevent medical malpractice. They are also used in the funding rules established between health organizations and the state or the health organizations and the insurance companies. Most of the times, the classification process is a manual and time-consuming process. In the case of the epilepsy, it often demands the realization of expensive complementary exams, e.g., electroencephalogram (EEG).

Hence, there is the opportunity to develop a new process, to reduce the time and effort to find and classify a correct diagnosis for each patient. To guarantee the adoption of this process by the medical community, the rationale behind each classification should be understandable by the physician.

However, developing a process to support ICD-9 diagnosis classification based on existing health records can be troublesome. Among other challenges it is possible to pinpoint that the information, usually free text, is not always structured in the same way, making it difficult to extract important and relevant knowledge. In fact, each physician usually has his own approach of describing events or symptoms, depending on their previous learning experiences and medical practices. On the other hand, the medical field has a specific language, which usually demands additional tools to interpret the designated terms and symptoms and to get semantic information from medical records. Additionally, mapping relevant features to the correct standard code classification can be difficult. Proper ICD-9 coding requires an understanding of how ICD-9 codes are used, how to use the ICD-9 manual, and the importance accuracy in ICD-9 coding.

The present work focuses on the ICD-9 classification of the epileptic diagnoses based on health records written in Portuguese for children under 16 years old. The proposed approach uses real electronic health records as source, including the preprocessing steps that use Natural Language Processing (NLP), followed by the definition of interpretable models that can be used in real diagnostic scenarios.

Next section introduces the concepts related with the approach, namely epilepsy, ICD-9 standard codes, health records, and text mining. Related work is discussed in Section 3 introducing some relevant projects on this field. Section 4 presents the proposed approach, discussing the general approach on classified medical records. Experiment setup and results are explained in Section 5 considering the case study scenario, results and discussion. Finally, in the last section, we will conclude and device new research steps.
2. Concepts

2.1. Epilepsy

Epilepsy consists of a number of recurrent and unpredictable seizures that occur through time [3]. A seizure is a manifestation of brain electrical discharges that will cause symptoms according to the specific location they occur in the brain. Due to these electrical discharges, the brain cannot perform normal tasks causing, e.g., seizures, language disturbances, hallucinations and absences. Not all seizures are epileptic; an alarm is set only when the seizures occur often (at least two times), not being provoked by alcohol, drogues, poisoning or other diagnosed diseases [3].

Epilepsy can be classified in different ways according, among others, the reason of the first seizure, patient observation during the episode, original location in the brain or the events that started the seizure. Additionally, there are other classifications that can help in epilepsy diagnosis. For example, it is possible to classify a seizure in different ways, but usually they are classified as partial, generalized or unknown [4]. Partial seizures are an electric discharge that was originated in a specific area of the brain. Generally these seizures begin in a specific location but can spread to other locations developing other symptoms. Generalized seizures are a chemical instability in both sides of the brain. Unknown seizures or idiopathic epilepsy is a classification where it is not possible to know the cause of the disease. Exams, such as electroencephalogram (EEG), computerized tomography, and physical exams can help in this classification. Moreover, medical or family history can help identifying previous types of seizures to support diagnoses.

2.2. Standard Codes Classification

Nosology is the systematic classification of diseases. In the twentieth century, when medical insurance programs made payers other than patients responsible for medical care, nosology became a matter of great interest to those public and private payers [5]. The most commonly used nosologies include ICD9, ICD10 or Systematized Nomenclature Of Medicine Clinical Terms (SNOMED-CT) [6]. These nosologies uniquely identify every diagnosis, description of symptoms and cause of death attributed to human beings. The use of these codes has expanded from classifying morbidity and mortality information for statistical purposes to diverse sets of applications, including administration, epidemiology, and health services research. The standardized codes improve consistency among physicians in recording patient symptoms and diagnoses.

It is possible to map one standard to other using tools like Unified Medical Language System (UMLS), also these systems can provide help with medical vocabulary, relations, syntax and morphology.

Epilepsy ICD-9 classification is in the group of “other disorders of the central nervous system (340-349)” [7]. The classification number assigned is the 345 and there are 10 classification possibilities: 1) 345.0 Generalized nonconvulsive epilepsy; 2) 345.1 Generalized convulsive epilepsy; 3) 345.2 Petit mal status; 4) 345.3 Grand mal status; 5) 345.4 Localization-related (focal) (partial) epilepsy and epileptic syndromes with complex partial seizures; 6) 345.5 Localization-related (focal) (partial) epilepsy and epileptic syndromes with simple partial seizures; 7) 345.6 Infantile spasms; 8) 345.7 Epilepsia partialis continua; 9) 345.8 Other forms of epilepsy and recurrent seizures; 10) 345.9 Epilepsy, unspecified. Other possible seizure-related codes, such as 779.0 Convulsions in newborn, 780.02 Transient alteration of awareness, 780.2 Syncope and collapse, 780.31 Febrile convulsions, and 780.39 Other convulsions and procedure codes are also taken into account.

In this work, given the data available in the field, we focus on efforts in the 345.1, 345.4 and 345.5 classifications.
2.3. Medical Information Systems

There are different types of medical information structures, like hospital information systems (HIS), electronic medical records (EMR) or electronic health records (EHR). HIS is a system that can manage medical, administrative, financial and legal aspects of a hospital [8]. EHR are a collection of medical records from individual patients or from a population. These records allow tracking patients and offer decision support mechanisms to access patient information across facilities of an institution [9].

Electronic medical records (EMR) includes patient demographics, summaries, medical history and lab tests, which are important for physicians to know the needs and take care of patients [10].

However, these data can be either structured or unstructured. Most of the times, data is in free text, with a specific semantic depending on each medical school or hospital and can have a very particular language where it is necessary additional techniques or vocabularies to make medical terms understandable to a machine. This makes the perception of content more difficult to understand requiring more effort and time to extract and classify.

2.4. Data and Text Mining

Data mining is the process of understanding and discovering patterns in large data sets to retrieve important knowledge [11]. This knowledge helps finding patterns improving, among others, the process of classification of diseases, saving time and money.

There are different techniques that can be used in the process of data mining, such as, association, classification, clustering, and prediction [12]. Data mining also uses machine learning methods, generally evolved from artificial intelligence that comprise algorithms to learn from data, constructing models that can classify information that was not previously learned. There are different learning strategies that can be pursued, namely, supervised learning, unsupervised learning and semi-supervised learning [13]. Supervised learning is the process of constructing models based on input-output examples given by a supervisor. Unsupervised learning aims at classifying entities on information without knowing the correct result. Semi-supervised learning is a learning process where only partial information is given to achieve a correct output.

When the focus of data mining includes text as input, there is a specialization area, text mining, which emphasis on the extraction of information from texts [14]. The texts can be in a structured or unstructured (harder to extract important information) format.

Text mining is a complex process because it requires the study of the frequency of words, word classification, understanding the meaning of each word, lexical and syntactic analysis and it is necessary to take into account what is really needed from the text. Moreover, it can become more complex since problems usually exhibit large dimensionality (number of features in input variables that must be taken into account).

Having such a specific input format makes it necessary to perform specific processing actions. First, it is safer to execute a pre-processing spell checking, stopword removal and document structure analysis [15]. Tokenization [16], splitting the text into words, phrases or other elements and stemming [17], which consists in identify words with a small syntactic variation (e.g. wait, waiting, etc.), are also needed.

Applying NLP techniques approaches include negation handling and name entity recognition [18] that is used to classify entities by analyzing words, classes, similar terminology, and abbreviations. Finally, word sense disambiguation is yet another technique that can be used in preprocessing to understand the meaning of each term based on the context [16]. After pre-processing, it is possible to apply text mining techniques.

There are different techniques in text mining, some adapted from data mining, such as, text summarization, information retrieval, and clustering. Text summarization captures the most important points in a text to create a summary. Information (document) retrieval allows locating and extracting information through user queries.
Clustering is an unsupervised technique that discovers groups of similar cases [19]. There are other characteristics for a faster and better text extraction like vector space model [20]. Vector space model [12] is a technique which represents documents or searches by vectors and tries to find similarities between them. These vectors have the necessary keywords extracted from the respective documents. It is also possible to make use of ontologies to achieve the classification faster and simpler. Ontologies are, generically, a list of concepts organized with classes, subclasses, properties, attributes and instances that can be useful to retrieve identifiers in documents to describe words and their relations [21].

3. Related Work

In this section we discuss works presenting epilepsy approaches with text mining and standard codes. There are some applications using text mining based on standard codes, like Computer Assisted Medical Information Resources Navigation & Diagnosis Aid Based on Data Marts & Data Mining (CAIRM-DAMM), which is a project applied to Areteion University Hospital in Greece with the objective of managing documents, multimedia documents retrieval, classify diagnoses based on ICD-9 and Data Mart. This project can also store medical information e.g., multimedia or texts, organize and retrieve documents based in Natural Language Queries (NQL) [22]. NQL is a system that interprets human language allowing queries based on uncontrolled terms, e.g., keywords that are not known, that can be present in documents. These keywords are classified as an entity, e.g., “diagnostic” or “person”. Furthermore, a ranked list is used to retrieve the correct classification according to the uncontrolled terms and their relationships present in the document. Each document is represented by a vector of existent uncontrolled terms, where ICD-9 diagnoses are proposed using classification rules.

Another example is a study to help professionals assigning ICD codes of the Swiss University Hospital of Geneva, using ICD-10, French thesaurus to identify the words, and information from the institution [23]. This project uses classification tasks based on ranking and multiclass instead of binary classification achieving better precision. First, data is pre-processed using stopword removal, negation handling, stemming, quality restoration (misspellings, diacritics), format normalization, and data acquisition. Then, a set of supervised learners is used and data-poor categorizer to assign unknown diagnoses that are not represented in the knowledge based of the institution.

There are other works using text mining on different fields including epilepsy, for example the work for the Danish psychiatric hospital [24] that can extract information by gathering phenotypic descriptions of patients from medical records and classify based on ICD 10 ontology to obtain patient stratification and disease co-occurrence statistics.

A different study was conducted in several health maintenance organizations served by Kelsey-Seybold clinics in Houston, to develop an algorithm that could detect epilepsy cases, based on combinations of diagnosis, diagnostic procedures, and used medication on electronic medical records, according to the standard code ICD-9 [25]. This study focused on building an algorithm that could maximize the sensibility and specificity, to increase the positive predictive value, lowering the false positive cases.

Other possible example is the investigation on epilepsy in children who enrolled in any type of school with attention-deficit or hyperactivity disorder. This study analyze the incidence and characteristics of epilepsy among population, based on electronic medical records. Characteristics of seizures, tests, and treatments were considered, to create a diagnosis and initiate treatment for attention-deficit or hyperactivity disorder in children with epilepsy [26].
Our work focuses ICD-9 classification of the epileptic diagnoses based on health records written in Portuguese. It is ontology-based using a white-box approach, i.e. physicians can understand why and how the system classified a disease of a patient, showing symptoms and rules for a classification.

4. Proposed Approach

The proposed process classifies epileptic diagnoses into a standard code. It uses a preprocessing step, where the documents are first cleaned, to identify and extract information, using a spell checker, replacing acronyms, removing duplicated characters and applying grammar rules. At that moment, a tokenizer is used to classify words, sentences and punctuation marks. A tagger classifies each word grammatically and a stemmer identifies words with small syntactic variations. Entity recognition and ontology tools were required to classify words, sentences and punctuation marks. A tagger classifies each word grammatically and a stemmer identifies words with small syntactic variations. Entity recognition and ontology tools were required to classify words, sentences and punctuation marks. At this point, different relations and rules between words are provided, e.g. finding several episodes, epilepsy history, family epilepsy history, loss awareness, or irregular movements. With these rules, entities are classified to give information to be used by learning algorithms, to create a model that can help classifying new words or rules, which can appear in the text. An ontology was developed to provide additional knowledge about some words, making it possible to find expressions that could help classifying the epilepsy and the different types of seizures. Then, with the help of rules, it is possible to create relations between entities for the learning process, where it is identified the correspondent ICD-9 code according to the relevant features.

Figure 1 shows the general proposed approach to classify epilepsies according with the ICD-9. If a patient has epilepsy, electronic medical records are analyzed. Extraction process of relevant information though a preprocessing step is reused, then, with the help of rules, it is possible to extract the most important features, mapping them into a standard code. There are some features that are important to extract in order to distinguish generalized from partial seizures, e.g., tonic and clonic movements, absence, loss awareness, psychomotor, somatomotor symptoms, convulsive seizures, and age. A machine learning algorithm classifies the best results against a corresponding standard code. With this classification, the correct procedures and prescriptions are shown to the physician.

![Figure 1 - Process of epileptic seizure type classification](image)

The system can learn through the physician with new clinical cases. It can learn new symptoms or adapt the way of classification according to the feedback of the physician. This will be possible by adjusting the classification that already exists or simply add new meaning of words to an existent category.
5. Experimental Setup and Results

5.1. Frameworks

There are several tools that can help preprocessing, such as Weka\(^1\), Rapid Miner\(^2\), R\(^3\) and General Architecture for Text Engineering\(^4\).

GATE is considered one of the best tools for language processing and information extraction for text mining [27]. This tool allows using ontologies, tokenizer and machine learning to classify information. GATE is also one of the most used applications in the medical field and was therefore elected in our approach. GATE has, however, some restrictions. Some plugins can only classify English thus, to use Portuguese texts, it was necessary to get other tools that could classify Portuguese language in a more complete way for the language dependent plugins. Freeling\(^5\) a tool that supports Tagging, Stemming and Entity Recognition in several languages, including Portuguese, was selected.

An integration tool was created to join the output of the different tools and, with the help of ontologies, create annotations that support the development of rules to find the relevant characteristics that will help the classification of the epilepsy according ICD-9 codes. The created ontology was based in Unified Medical Language System (UMLS)\(^6\), which can give knowledge about Portuguese words or expressions, such as anatomy or events, necessary for classification. The rules were implemented with Java Annotation Patterns Engine (JAPE). This will identify on text words or patterns from annotations in order to select the relevant features. For example if it is found an annotation of “movement” next to an “abnormal” or “involuntary” and before any punctuation mark or conjunction, then it is identified as “involuntary movement”. In the end, annotations are created for all features, to provide information used by the GATE machine learning. Since GATE does not support numeric features, it is necessary to use another tool for machine learning, such as Weka Therefore, an Attribute-Relation File Format (ARFF) was created with JAPE to export the results obtained in pre-processing to Weka, as we can see in Figure 2.

Figure 2 – Presented architecture

---

1 http://www.cs.waikato.ac.nz/ml/weka
2 http://rapid-i.com
3 http://www.rdatamining.com
4 http://gate.ac.uk
5 http://nlp.lsi.upc.edu/freeling
6 http://www.nlm.nih.gov/research/umls
5.2. Dataset – Real case study

The process was tested with a real dataset, constructed with real anonymous patient records, provided by a hospital. The records contain epilepsy diagnoses. For each of these records the final diagnosis was defined, as well as the main features that led to that diagnosis. With this information it was possible to determine relevant features that could lead to a possible epileptic seizure type. Then, we build rules in JAPE to find these features and mark them as annotations to be understandable by GATE Machine Learning. With this classification it was possible to check if a word appeared or not on the records and if was mention as negative symptom, such as “didn’t had seizures”. A numeric classification was defined, where “-1” represented a negative symptom, “1” if a symptom was encountered and “0” if was not mentioned. Subsequently, intensity is verified, to give more detail to each feature. If a patient had intense loss awareness, this feature would have more importance, then a simple mention of loss awareness on text. In addition, was described a nominal classification for training the final class, i.e. was annotated the seizure type according to the standard code of the final diagnosis, e.g. “345.5” that means “Localization-related (focal) (partial) epilepsy and epileptic syndromes with simple partial seizures” according to the ICD-9.

Table 1 shows the set of seizure types that were found on the electronic medical records provided.

<table>
<thead>
<tr>
<th>Seizure Type</th>
<th>Frequency on dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex focal seizure</td>
<td>10</td>
</tr>
<tr>
<td>Simple focal seizure</td>
<td>3</td>
</tr>
<tr>
<td>Generalized convulsive epilepsy</td>
<td>6</td>
</tr>
</tbody>
</table>

5.3. Evaluation Metrics

To tackle a multiclass problem as the one present different approaches can be followed. We will divide text mining problem into several two-class problems, using a one-vs-all approach. To evaluate the decision task, we first define possible outcomes of the classification: true positive (a), false positive (b), false negative (c), and true negative (d). Several measures have been defined based on these values, e.g., error rate ((b+c)/(a+b+c+d)), recall (R=a/(a+c)), and precision (P=a/(a+b)), as well as combined measures, such as, the F1 measure, combining recall and precision in a single score: F1=2*P*R/(P+R). F1 is one of the best-suited measures for text classification, since it deals well with unbalanced scenarios, common in text classification.

Having several classifiers some form of averaging has to be used to find total criteria values. There are two types of averaging: micro-averaging and macro-averaging. In micro-averaging, performance tables for each of the categories are added, and the criteria are computed. In macro-averaging, performance measures are computed separately for each category and the mean of the resulting performance is taken. The results presented in this paper use macro-averaging.

5.4. Learning and Results

Machine Learning has different processes of deducting models (functions) from information, which can be used to map new documents. A multiclass classification using one-vs-all with K-Nearest Neighbor (KNN) algorithm was chosen, assigning the class most common in the K nearest neighbors.

Table 2 shows the difference between the initial results with simple focal seizures and without them, with a value for K=3, using 3-fold and cross-validation. Cross-validation was considered given the number of patient
records that were available. Tests were carried out with 19 records, which while undoubtedly low, still make it possible to perform some analysis of the obtained results.

Analyzing these results, it is possible to conclude that records of simple focal seizure are extremely scarce for the learning procedure to have results, making it impossible to obtain any true positive instance or a valid F1 measure. Therefore, other tests were carried out, removing the examples for simple focal seizure from the dataset. Results presented in Table 3 present a slight improvement over initial results, showing that if the dataset is richer regarding the quality of representation of each epilepsy gains can be attained.

These results are only preliminary with the risk of over fitting training models, which can occur when handling a complex model with more features than examples.

Table 2 – Initial results on seizure type classification

<table>
<thead>
<tr>
<th>Seizure Type</th>
<th>FP</th>
<th>FN</th>
<th>TP</th>
<th>TN</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex focal seizure</td>
<td>1</td>
<td>5</td>
<td>9</td>
<td>4</td>
<td>73%</td>
</tr>
<tr>
<td>Generalized convulsive epilepsy</td>
<td>4</td>
<td>10</td>
<td>2</td>
<td>3</td>
<td>62.2%</td>
</tr>
<tr>
<td>Simple focal seizure</td>
<td>3</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Further analyzing the performance results, we can see that the number of false negatives has been significantly reduced, specially in the complex focal seizure case (from 5 to 3), which is extremely relevant in medical settings, such as epilepsy diagnosis. These incorrect false negative instances are usually alarming, i.e. these errors are important to avoid because getting a treatment to control seizures, in most people stops the seizures from occurring.

Results so far are encouraging with an F1 average measure of 71.05%; however more medical records and different seizure types will be needed to get more confident results.

Table 3 – Preliminary results on seizure type classification

<table>
<thead>
<tr>
<th>Seizure Type</th>
<th>FP</th>
<th>FN</th>
<th>TP</th>
<th>TN</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex focal seizure</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>3</td>
<td>74%</td>
</tr>
<tr>
<td>Generalized convulsive epilepsy</td>
<td>3</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>68.1%</td>
</tr>
<tr>
<td>Weighted Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>71.05%</td>
</tr>
</tbody>
</table>

6. Conclusions and Future Work

This paper proposes a process to support pediatric physicians’ decisions in the context of a real scenario. The proposed approach uses real health records as source, preprocessing steps that use NLP followed by the definition of interpretable models that can be used in existent scenarios. Results are still preliminary, which means that a substantial volume of training data, more records with other seizure type’s classification, and extraction of more relevant features is needed to gain more precision in the process. Physicians are still apprehensive with this kind of technology because of the low level of accuracy. Therefore it is important continuing to follow a white box approach in this work. Future work will expand the real dataset and deal with dynamic issues, as new patient records appear every day.
Acknowledgements

The authors would like to thank to the Leiria-Pombal regional hospital (Hospital Santo André) that provided real anonymous patient records, which were crucial for this work.

This work has been partially supported by the Portuguese Foundation for Science and Technology under project grant[s] PEst-C/ EEI/UI308/2011.

References