Journal of Universal Computer Science, vol. 20, no. 6 (2014), 907-923
submitted: 1/7/13, accepted: 14/2/14, appeared: 1/6/14 © J.UCS

Decision Support System to Diagnosis and Classification of Epilepsy in Children

Rui Rijo
(INESCC - Institute for Systems and Computers Engineering at Coimbra - Portugal
School of Technology and Management, Polytechnic Institute of Leiria, Portugal
rui.rijo@ipleiria.pt)

Catarina Silva
(CISUC – Centre for Informatics and Systems of the University of Coimbra - Portugal
School of Technology and Management, Polytechnic Institute of Leiria, Portugal
catarina@ {dei.uc.pt, ipleiria.pt})

Luís Pereira
(School of Technology and Management, Polytechnic Institute of Leiria, Portugal
luispereira268@gmail.com)

Dulce Gonçalves
(School of Technology and Management, Polytechnic Institute of Leiria, Portugal
dulce.goncalves@ipleiria.pt)

Margarida Agostinho
(Centro Hospitalar de Leiria-Pombal, Hospital Santo André, Leiria, Portugal
gi@gijoeha.net)

Abstract: Clinical decision support systems play an important role in organizations. They have a tight relation with the information systems. Our goal is to develop a system to support the diagnosis and the classification of epilepsy in children. Around 50 million people in the world have epilepsy. Epilepsy diagnosis can be an extremely complex process, demanding considerable time and effort from physicians and healthcare infrastructures. Exams such as electroencephalograms and magnetic resonances are often used to create a more accurate diagnosis in a short amount of time. After the diagnosis process, physicians classify epilepsy according to the International Classification of Diseases, ninth revision (ICD-9). Physicians need to classify each specific type of epilepsy based on different data, e.g., types of seizures, events and exams' results. The classification process is time consuming and, in some cases, demands for complementary exams. This work presents a text mining approach to support medical decisions relating to epilepsy diagnosis and ICD-9-based classification in children. We put forward a text mining approach using electronically processed medical records, and apply the K-Nearest Neighbor technique as a white-box multiclass classifier approach to classify each instance, mapping it to the corresponding ICD-9-based standard code. Results on real medical records suggest that the proposed framework shows good performance and clear interpretations, albeit the reduced volume of available training data. To overcome this hurdle, in this work we also propose and explore ways of expanding the dataset.

Keywords: epilepsy, diagnosis, clinical decision support systems, medical information systems, electronic medical records, ICD codes, data mining, text mining, machine learning
Categories: H.3.1, H.4.2

1 Introduction

Clinical decision support systems (CDSS) are computer systems designed to impact clinician decision making about individual patients at the point in time that these decisions are made [Berner, 07]. These systems may have a great impact in healthcare organizations by reducing the medical error, improving services, and support also better management decisions. They face project management challenges. These challenges need useful project management practices [Fernandes, 13]. This work focused in the development of a clinical decision support system to support diagnosis and classification of epilepsy in children.

According to the World Health Organization (WHO), epilepsy is the second most common neurological disease, affecting each year 50 million people around the world [WHO, 12], and 70.000 people just in Portugal (from a 10 million population) [LPCE, 13].

The process of identifying and classifying epilepsy is complex, demanding considerable time and effort [Brown, 00]. Several characteristics, symptoms and exams must be considered 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 to allow patients for a normal and productive quotidian. Moreover, this difficulty becomes intensified in children, since it requires the analysis of different causes such as genetic, structural or metabolic-based ones. 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 [Fogoros, 13]. These diagnoses must be classified according to 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 of America.

ICD-9 codes are used to describe a patient's diagnosis including symptoms, diseases or disorders. An ICD-9 code contributes for a common understanding and interpretation of a diagnosis. It is important a correct ICD-9 classification for the quality of patient care and to prevent medical malpractice. ICDs are also used in the funding rules established between health organizations and the state/insurance companies. However, quite often, this is a manual time-consuming classification process carried out by physicians.

In the case of epilepsy, the diagnosis often demands for expensive complementary exams, such as an electroencephalogram (EEG). Hence, there is the opportunity to develop new processes, to reduce time and effort to determine a correct diagnosis for each patient and its ICD-9 classification. To guarantee the adoption of such processes by the medical community, the rationale behind each classification should be understandable by the physician.

However, developing a process to support ICD-9 diagnosis and classification using existing health records can be troublesome. Among other challenges, it is possible to pinpoint that the information, usually written in 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 his 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 and the importance of accuracy in ICD-9 coding.

The present work focuses on the diagnosis and 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, and includes the pre-processing steps that use Natural Language Processing (NLP), followed by the definition of interpretable models that can be used in real diagnostic scenarios. The research was done by a multidisciplinary team and in collaboration with a paediatric service of a hospital.

Next section introduces the concepts related with this approach, namely epilepsy, ICD-9 standard codes, health records, data and text mining. Related work is discussed in Section 3, introducing relevant projects on this research area. Section 4 presents the proposed approach to achieve a diagnosis and to its further classification. Experimental setup and results are detailed in Section 5, considering the case study scenario, its results and discussion. Finally, in Section 6 we present the conclusions and identify new research steps.

2 Concepts

In this section we introduce the main concepts needed to provide context to our work. We begin with the concept of clinical decision support system, medical information systems and epilepsy, followed by the corresponding standard codes for its classification. We conclude by providing a short insight on data and text mining.

2.1 Clinical Decision Support Systems

Clinical decision support systems (CDSS) are designed to assist physicians or other professionals making more informed clinical decisions, helping in the diagnosis, and analysis [Romano, 11]. Furthermore, a decision support system can support the control of costs, among others, monitoring medication orders, managing clinical complexity, performing a preventive care, and supporting administrative tasks. These systems enable to reduce medical error, medication error and adverse drug events, saving time and money [Jaspers, 10].

CDSS systems may be described in terms of five things that they do “provide the right information, to the right person, in the right format, through the right channel, at the right point in the workflow to improve health and health care decisions and outcomes” [Osheroff, 04]. Such systems do not themselves perform clinical decision making; they provide relevant knowledge and analyses that enable the ultimate decision makers – clinicians, patients, and health care organizations – to develop more informed judgments [Musen, 14]. Clinical decision support systems manage information from medical information’s systems, presented in the next subsection.

2.2 Medical Information Systems

There are different types of medical information systems, such as the ones provided by hospital information systems (HIS), electronic health records (EHR) or electronic medical records (EMR). HIS are systems that can manage medical, administrative, financial and legal aspects of a hospital [Tsumoto, 11]. 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 [Hoerbst, 10]. EMR include patient demographics, summaries, medical history and laboratory tests [Ludwick, 08].

However, these data sources can be either structured or unstructured. Quite often, data come in free text, with specific semantics depending on each medical school or hospital, and can have a very particular language where additional techniques or vocabularies are necessary to make medical terms understandable to a machine. This makes the perception of content more difficult to accomplish, requiring more effort and time to extract and classify [Caballero, 12], making this one of the challenges to tackle in this work.

2.3 Epilepsy

Epilepsy consists of a number of recurrent and unpredictable seizures that occur through time [Engel, 12]. 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 [Engel, 12].

Epilepsy can be classified in different ways according to, 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 is other information 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 [Berg, 10]. 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 determine the cause of the disease. Exams, such as electroencephalogram (EEG), computerized tomography (CT), and physical exams can help with this classification. Moreover, medical or family histories are also relevant in the identification of previous types of seizures to support diagnoses.

2.4 Standard Codes Classification

Nosology is the systematic classification of diseases. In the 20th 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 [Armstrong, 11]. The most commonly used nosologies include the International Classification of Diseases with different revisons: ICD-9, ICD-10 and the

Systematized Nomenclature Of Medicine - Clinical Terms (SNOMED-CT) [Coonan, 04]. 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 another using tools like the Unified Medical Language System (UMLS) as these systems provide medical vocabulary, relations, syntax, and morphology.

Epilepsy ICD-9 classification is in the group of “other disorders of the central nervous system (340-349)” [ICD9Data.com, 13]. The classification number assigned is the 345 and there are 10 classification possibilities as shown by Table 1.

ICD-9 classification number	**Designation**
345.0	Generalized nonconvulsive epilepsy
345.1	Generalized convulsive epilepsy
345.2	Petit mal status
345.3	Grand mal status
345.4	Localization-related (focal) (partial) epilepsy and epileptic syndromes with complex partial seizures
345.5	Localization-related (focal) (partial) epilepsy and epileptic syndromes with simple partial seizures
345.6	Infantile spasms
345.7	Epilepsia partialis continua
345.8	Other forms of epilepsy and recurrent seizures
345.9	Epilepsy, unspecified

Table 1: ICD-9 classification of epilepsy

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”, 780.39 “Other convulsions”, exist in the ICD9 classification of epilepsy. In this work we focus our efforts in the 345.1, 345.4 and 345.5 classification codes, given the data available from our real-world case study.

2.5 Data and Text Mining

Data mining is the process of understanding and discovering patterns in large data sets to retrieve important knowledge [Fayyad, 96]. This knowledge helps finding patterns improving, among others, the process of classification of diseases, whilst saving time and money.

There are different techniques that can be used in the process of data mining, such as, association, classification, clustering, or prediction [Tan, 06]. Data mining often makes use of machine learning methods, generally evolved from artificial intelligence, that comprise algorithms to learn from data, constructing models that can classify cases that were not previously known. There are different learning strategies that can be pursued, namely, supervised learning, unsupervised learning and semi-supervised learning [Chaovalit, 05]. 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, by grouping similar inputs. 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 is centred on the extraction of information from texts [Hearst, 99]. Those texts can be in a structured or unstructured format. The latter ones imply more challenges to extract meaningful information.

Text mining is a complex process since it requires the study of the frequency of words, word classification, understanding the meaning of each word, and lexical and syntactic analysis. It is also necessary to take into account what is really needed from the text. Additionally, 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 [Hammouda, 04]. Tokenization [Witten, 03], i.e., splitting the text into words, phrases or other elements and stemming [Feldman, 98], which consists in identifing words with a small syntactic variation, e.g. “wait” and “waiting”, are also needed.

Applying NLP techniques include negation handling and name entity recognition [Krallinger, 05] that is used to classify entities by analysing words, classes, similar terminology, and abbreviations. Finally, word sense disambiguation is yet another technique that can be used in pre-processing to understand the meaning of each term based on the context [Witten, 03]. After pre-processing, it is possible to apply text mining techniques.

There are different techniques in text mining, some adapted from data mining, namely, 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 [Tseng, 07]. Another technique for faster and better text extraction is the vector space model [Han, 06]. Vector space model represents documents or searches by vectors and tries to find similarities between them [Tan, 06]. 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 within classes, subclasses, properties, attributes and instances that can be useful to retrieve identifiers in documents to describe words and their relations [Wongthongtham, 09].

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 (CAIRN-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 (NLQ) [Karanikolas, 03]. NLQ 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.

Ruch et al. [Ruch, 07] help clinical professionals assigning ICD-10 codes in the Swiss University Hospital of Geneva, making use of a French thesaurus to identify the words and also information from the institution. This project uses classification tasks based on ranking and multiclass classification, instead of binary classification, achieving better precision. First, data are 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.

Roque et al. [Roque, 11] conducted a research work for a Danish psychiatric hospital to extract information by gathering phenotypic descriptions of patients from medical records and classifying them based on an 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 [Holden, 05] to develop an algorithm that could detect epilepsy cases, based on combinations of diagnoses, diagnostic procedures, and used medication on electronic medical records, according to the standard code ICD-9. 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.

Davis et al. [Davis, 10] developed their work on epilepsy in children who enrolled in any type of school with attention-deficit or hyperactivity disorder. This study analyses the incidence and characteristics of epilepsy among population, based on electronic medical records. Characteristics of seizures, tests, and treatments were considered in order to create a diagnosis and initiate treatment for attention-deficit or hyperactivity disorder in children with epilepsy.

4 Proposed Approach

Our work focuses on clinical diagnosis and ICD-9 classification of the epileptic diagnoses based on health records written in Portuguese. It is ontology-based and can be easily used in other languages and uses white-box approaches, where physicians can understand why and how the system classifies a disease of a patient, showing symptoms and rules. To reach this goal, text mining is used to extract all the relevant information, in order to identify and classify entities, e.g. symptoms, to reach the objectives, and specifically to determine a diagnosis and its ICD-9 classification. This approach intends to support decisions, to reduce time, effort and medical error in a diagnosis, treatment, prescriptions and ICD-9 classification.

Figure 1 depicts the general proposed approach to process medical information, reach a diagnosis, and classify epilepsy. The most important input are the patients' medical records that include the information that physicians write about each patient. This input is received as plain text and is processed using text mining techniques.

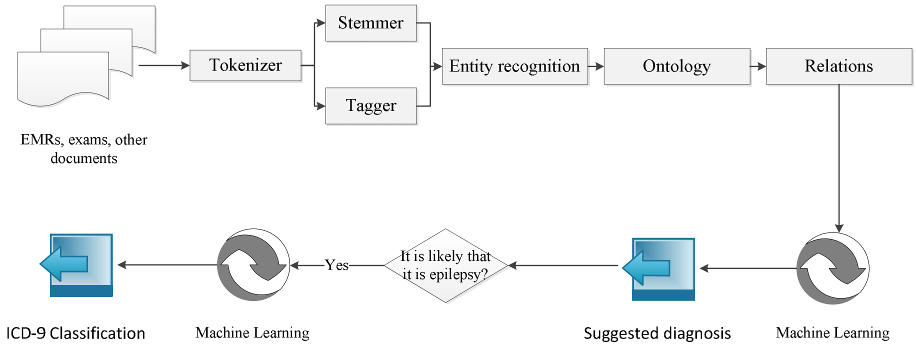

Figure 1: Proposed approach to process medical information

We start with a pre-processing step. In this step documents are first cleaned using a spell checker, replacing acronyms, removing duplicated characters and applying grammar rules. This helps to identify and extract relevant information.

Then, a tokenizer classifies words, sentences and punctuation marks. A tagger is then used to classify each word grammatically and a stemmer identifies words with small syntactic variations. Entity recognition and ontology tools categorize the entity for each word. An ontology was developed to provide additional knowledge about some words, making it possible to find expressions that could help classifying probable epilepsy.

At this moment, different relations and rules between words are provided, e.g. finding more than one seizure, epilepsy history, family epilepsy history, loss of awareness, or irregular movements. With these rules, entities are classified and the result is a dataset that can be used by learning algorithms to create a model that can classify new words or rules, which can appear in the text.

A supervised learning approach is then pursued, since there is a classification for each record, i.e., each one has a final diagnosis to build a model to classify future records. For each record classified as probable epilepsy, a new model, previously created also with a supervised learning, suggests the ICD-9 classification.

The system can also learn through the physician with new clinical cases. It is possible that the system learns new symptoms, or adapts the way of classification according to the physician. This will be possible by adjusting the classification that already exists or simply add new meaning of words to an existent category.

Cross Industry Standard Process (CRISP) is the methodology chosen for this research. This methodology has different steps, in concrete, business understanding, data understanding, data preparation, modeling, evaluation and deployment. Business understanding allows the insight of objects and requirements to be achieved. At that time, it is necessary to understand the information, knowledge that is important to remove and prepare the data in order to build models, that are evaluated and the best that fits the problem is chosen.

5 Experimental Setup and Results

5.1 Frameworks

We use the General Architecture for Text Engineering (http://gate.ac.uk/) framework for text mining. GATE is considered one of the best tools for language processing and information extraction for text mining [Ruch, 07]. It allows the use of 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. It has, however, some restrictions. Some plugins can only classify English. Therefore, 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 in GATE. For this purpose, Freeling (http://nlp.lsi.upc.edu/freeling/), a tool that supports Tagging, Stemming and Entity Recognition in several languages, including Portuguese, was selected. Freeling is used to preprocess Portuguese language, classifying the words grammatically, and to find some relevant entities through the tagger, stemmer and entity recognition features. The tool receives documents (in plain text) and identifies a possible classification for each word or expression.

Figure 2 presents a general view of the software architecture to support the proposed approach to process medical information, achieve a diagnosis and classify epilepsy.

An integration engine was created to join these classifications and to add annotations, with the help of ontologies. These annotations will support the development of rules to find relevant characteristics that, in turn, will classify the probability of a patient having epilepsy. The developed ontology is based in the Unified Medical Language System (UMLS) (http://www.nlm.nih.gov/ research/umls/), which offers knowledge about Portuguese words or expressions, such as anatomy or events, paramount for classification. The ontology was built in Protégé (http://protege.stanford.edu/), an open source ontology and knowledge-based framework. The rules were implemented with the Java Annotation Pattern Engine (JAPE), a plugin of GATE. With this module (Relations) capable of creating relations between entities, the relationships are established to be the learning base of supervised classification. The machine learning setting of GATE was also used to identify probable classifications and other rules on text that were not previously specified. This will identify words or patterns from annotations, on text, 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 that can be used by the GATE machine learning. Since GATE does not support numeric features, we used another tool for machine learning, namely Weka [Roque, 11]. Therefore, an Attribute-Relation File Format (ARFF) was created with JAPE, to export the results obtained in pre-processing in GATE to Weka (http://www.cs.waikato.ac.nz/ml/weka/).

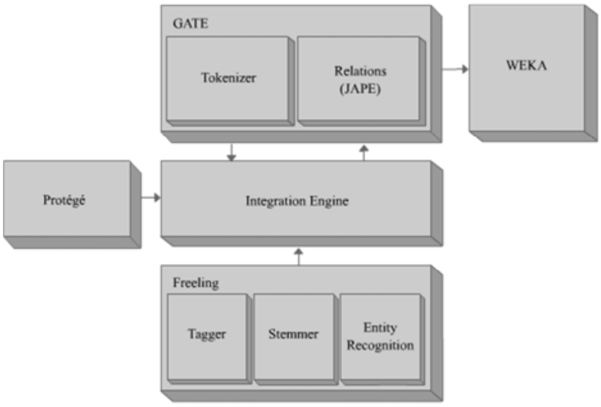

Figure 2: Software architecture to support the proposed approach

5.2 Dataset – Real-world case study

The process was tested with a real-world dataset, constructed with 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 built 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 it was mentioned as a 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 not mentioned. Subsequently, intensity was verified, to give more detail to each feature. If a patient had intense loss awareness, this feature would have more importance than a simple mention of loss awareness on text. In addition, a nominal classification for training the final class was put forward, i.e. the seizure type was annotated according to the standard code of the final diagnosis, e.g. "345.5", which means "Localization-related (focal) (partial) epilepsy and epileptic syndromes with simple partial seizures", according to ICD-9.

The procedure of gathering medical records was very time consuming and did not allowed for the collection of a large number of records or a wide diversity of symptoms and diagnoses. All medical records were extensive, in paper, and handwritten. The team spent a considerable amount time in the digitalization of those records and understanding the key characteristics of epilepsy in children. Therefore, only some types of diagnosis were discovered and an initial test with 19 complete medical records was carried out. Table 2 shows the set of seizure types that were found on the medical records provided.

		Frequency on dataset	**ICD-9 code**
Seizure Type	Complex focal seizure	10	345.4
	Simple focal seizure	3	345.5
	Generalized convulsive epilepsy	6	345.1

Table 2: Seizure type frequency on dataset

5.3 Evaluation Metrics

To tackle a multiclass problem as the present one, different approaches can be followed. We opted to divide the 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, TP as (a), False Positive, FP as (b), False Negative, FN as (c), and True Negative, TN as (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, which are quite 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 for the diagnosis

Machine Learning has different processes of deducting or inducing models (functions) from data, which can be used to map new documents. To achieve a diagnosis for each dataset record, a simple K Nearest Neighbor (KNN) algorithm was chosen, assigning the class most common among the K nearest neighbors. Table 3 shows the F1 results for different values for K and different numbers of cross-validation folds. Cross-validation was considered given the number of patient records that were available. Tests were carried out with 19 records. This number of records is undoubtedly low, but still it is possible to perform some analysis of the obtained results. From Table 3, we conclude that the best scenario is with K=1. These results are only preliminary with the risk of over fitting training models, which can occur when handling complex models with more features than examples.

The incorrect classifications are usually alarming mainly if they are false negatives, i.e. instances that were classified as not having epilepsy, but in fact are epilepsies. It is crucial to avoid these errors because getting a treatment to control seizures, in most people, stops the seizures from occurring. Results so far are encouraging. However, more medical records will be needed to get more confident results.

		Cross-validation (folds)			
		2	3	4	5
K-value	1	88.89%	88.89%	100%	100%
	2	77.78%	88.89%	100%	100%
	3	77.78%	66.67%	77.78%	77.78%

Table 3: F1 performance for KNN

5.5 Learning and Results for the ICD9 classification

In order to achieve an ICD-9 classification for each dataset record, a multiclass classification using *one vs. all* with KNN algorithm was chosen, assigning the class most common in the K nearest neighbors.

Table 4 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 acceptable 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 5 present a slight improvement over initial results, showing that gains can be attained providing a richer dataset, regarding the quality of representation of each type of epilepsy,.

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

Seizure Type	**FP**	**FN**	**TP**	**TN**	**F1**
Complex focal seizure	1	5	9	4	73.0%
Generalized convulsive epilepsy	4	10	2	3	62.2%
Simple focal seizure	3	15	0	1	N/A

Table 4: Initial results on seizure type classification

Further analyzing the performance results, we can see that the number of false negatives has been significantly reduced, especially in the complex focal seizure case (from 5 to 3), which is extremely relevant in medical settings. These incorrect false negative instances are usually alarming, i.e. these errors are important to avoid because getting the right 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.

Seizure Type	FP	FN	TP	TN	F1
Complex focal seizure	1	3	9	3	74.0%
Generalized convulsive epilepsy	3	8	3	2	68.1%
Weighted Average					71.05%

Table 5: Preliminary results on seizure type classification

5.6 Expanding the dataset

Due to the limited number of electronic medical records available and the irregular distribution of their characteristics, it was necessary to expand the dataset. We decided to use the crossover technique for this purpose, as it allows the combination of parts of the existing records, generating new records, when applied to the available medical records. For this, medical records were divided into approximately equal parts. These parts are the input to the crossover technique. Selecting randomly these parts we built new electronic medical records. Figure 3, shows the crossover technique over the records.

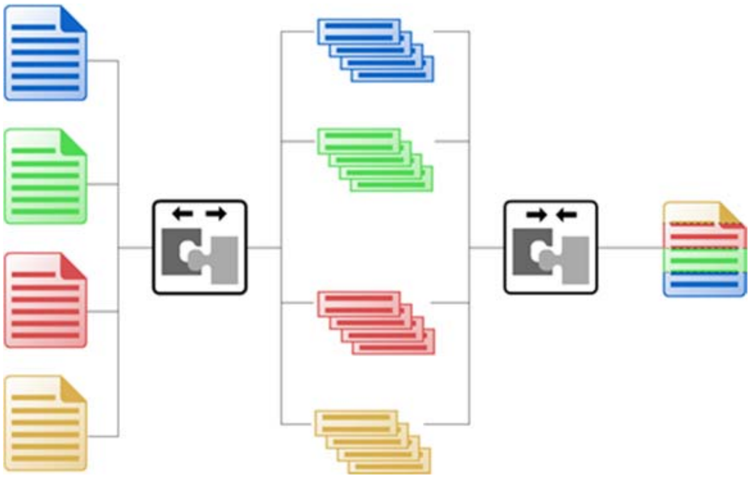

Figure 3: Crossover applied to medical records

As these new records may be illogical from the medical perspective, the analysis of the medical team was necessary to assure validation for the classification process of a probable epilepsy and ICD classification. After the process we got 91 records classified with a diagnosis of epilepsy and 22 records without epilepsy.

5.7 Learning and results from the expansion of the dataset applied to the diagnosis classification

We applied the expanded dataset to the diagnosis classification, using KNN (IBK in Weka), Fuzzy Unordered Rule Induction Algorithm (FURIA) [Hühn, 09] and Simple Classification & Regression Tree (Simple CART) [Cufoglu, 09]. FURIA learns fuzzy rules instead of conventional rules and unordered rule sets instead of rule lists. Moreover, to deal with uncovered examples, it makes use of an efficient rule stretching method. Simple CART is a classification technique that generates the binary decision tree. Since output is binary tree, it generates only two children. Table 6 presents the results for the classification of a probable diagnosis of epilepsy.

In the experiments we applied 20-fold cross validation for the FURIA and KNN (IBK in WEKA) and 16-fold cross for the CART algorithm in order to get a more realistic rating. Regarding KNN, we used K=3 for the first test and K=5 for the others, since new symptoms were added and it was suitable to use a larger neighborhood for classification.

		TP	**FP**	**FN**	**TP**	**Accuracy**	**F1**
Test 1	SimpleCART	2	20	2	89	80.53%	74.7%
	FURIA	4	18	4	87	80.50%	76.7%
	IBK (K=3)	6	16	5	86	81.40%	78.8%
Test 2	SimpleCART	1	21	3	88	78.76%	72.4%
	FURIA	3	19	4	87	79.65%	75.2%
	IBK (K=5)	4	18	5	86	79.65%	76.1%
Test 3	SimpleCART	1	21	3	88	78.76%	72.4%
	FURIA	2	20	3	88	79.60%	74.1%
	IBK (K=5)	4	18	4	87	80.53%	76.7%

Table 6: Results of final tests for classifying probable epilepsy

From the results, we conclude that the dataset is still unbalanced, i.e., there are records that correspond to real epilepsy but are classified as not being epilepsy (false negatives, FN). We can also verify that several true negatives cases (records without epilepsy) are classified as false positives (epilepsy diagnosis, but without the real existence of the disease).

Despite these aspects, it is still possible to say that these results are encouraging, with an F1 measure of around 78%.

6 Conclusions and Future Work

This paper proposes a process to support pediatric physicians' decisions in the context of a real-world scenario of epilepsy diagnosis. The proposed approach uses real health records as source, pre-processing steps that use NLP, followed by the definition of interpretable models that can be used in existent scenarios.

To suitably test the proposed framework, a real-world dataset was constructed using real anonymous health records. Moreover, a strategy for expanding the dataset was put forward, resulting in a six-times larger dataset.

Results so far show that the proposed framework shows good performance and clear interpretations, albeit the reduced volume of available training data.

Future work is foreseen in substantially enlarging the dataset, since having more records with other seizure type's classification, and providing the extraction of more relevant features will provide greater performance to the results. Additionally, it is important to continue to follow white-box approaches, since physicians are still apprehensive with this kind of technology because of its general low level of accuracy. We intend to apply FURIA and SimpleCART to the ICD-9 classification.

Finally, we intend also to classify the expanded dataset with ICD-9 and deal with dynamic issues associated with the online use of the framework.

Acknowledgements

The authors would like to thank to the pediatric service of Centro Hospitalar Leiria-Pombal, namely to Hospital Santo André, which provided real anonymous patient records and participated with a clinical team in this research.

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

References

[Armstrong, 11] Armstrong, D., Diagnosis and nosology in primary care. Sociology of Diagnosis, 2011. 73(6): p. 801–807.

[Berg, 10] Berg, A.T., et al., Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology. Epilepsia, 2010. 51(4): p. 676-685.

[Berner, 07] Berner, E. S. Clinical Decision Support Systems. Springer Science+ Business Media, LLC, 2007.

[Brown, 00] Brown, R.J. and M.R. Trimble, Dissociative psychopathology, non-epileptic seizures, and neurology. J Neurol Neurosurg Psychiatry, 2000. 69(3): p. 285-9.

[Caballero, 12] Caballero, I., Sánchez, L. E, Freitas, A., Fernández-Medina, E. HC+: Towards a Framework for Improving Processes in Health Organizations by Means of Security and Data Quality Management, Journal of Universal Computer Science, 2012. 18(12): p. 1703-1720.

[Chaovalit, 05] Chaovalit, P. and L. Zhou, Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches, in Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.

[Coonan, 04] Coonan, K.M., Medical informatics standards applicable to emergency department information systems: making sense of the jumble. Academic Emergency Medicine, 2004. 11(11): p. 1198-1205.

[Cufoglu, 09] Cufoglu, A.; Lohi, M.; Madani, K., A Comparative Study of Selected Classifiers with Classification Accuracy in User Profiling. Computer Science and Information Engineering, 2009 WRI World Congress, 2009. .3: p.708-712

[Davis, 10] Davis, S.M., et al., Epilepsy in Children with ADHD: A Population-Based Study. Pediatric Neurology, 2010. 42(5): p. 325-330.

[Engel, 12] Engel, J., Seizures and Epilepsy. 2 ed2012: Oxford University.

[Fayyad, 96] Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. From Data Mining to Knowledge Discovery in Databases. AI Magazine, 1996. 17: p. 37-54.

[Feldman, 98] Feldman, R., Fresko, Y. Kinar, Y. Lindell, O. Liphstat, et al., Text mining at the term level, Proc. Of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98), Nantes, France, September 23-26, 1998.

[Fernandes, 13] Fernandes, G., Ward, S., and Araújo, M., Identifying useful project management pratices: A mixed methodology approach. International Journal of Information Systems and Project Management, 2013, 1(4): p.5-21

[Fogoros, 13] Fogoros, R.N. The Misdiagnosis of Epilepsy. 2009 May 20th 2013. Available from: http://heartdisease.about.com/b/2009/08/07/the-misdiagnosis-of-epilepsy.htm.

[Han, 06] Han, J. and M. Kamber, Data Mining Concept and Techniques 2ed, 2006, San Francisco: Morgan Kaufmann.

[Hearst, 99] Hearst, M.A., Untangling text data mining, In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL '99). Association for Computational Linguistics, Stroudsburg, PA, USA, 1999. p. 3-10.

[Hammouda, 04] Hammouda, K.M. and M.S. Kamel, Efficient Phrase-Based Document Indexing for Web Document Clustering, in IEEE Transactions on Knowledge and Data Engineering, 2004. p. 1279-1296.

[Hoerbst, 10] Hoerbst, A. and E. Ammenwerth, Electronic Health Records. Methods of Information in Medicine, 2010.

[Holden, 05] Holden, E.W., et al., Developing a Computer Algorithm to Identify Epilepsy Cases in Managed Care Organizations. Disease Management, 2005. 8(1): p. 1-14.

[Hühn, 09] Jens Hühn and Eyke Hüllermeier, FURIA: an algorithm for unordered fuzzy rule induction. Data Mining and Knowledge Discovery, 2009. 19(3): p. 293-319.

[Jaspers, 10] Jaspers, M. W. M., Smeulers, M., Vermeulen, H., & Peute, L. W. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. Journal of the American Medical Informatics Association, 2010. 18: p. 327-334.

[Karanikolas, 03] Karanikolas, N.N. and C. Skourlas, Shifting from legacy systems to a data mart and computer assited information resources navigation framework. CEIS 2003 - Databases And Information Systems Integration, 2003: p. 300-305.

[Krallinger, 05] Krallinger, M., Ramon, A-A., and Valencia. A., Text mining approaches in molecular biology and biomedicine. Drug discovery today: biosilico, 2005. 10(6): p.439-445.

[LPCE, 13] LPCE – Liga Portuguesa Contra a Epilepsia. Generalidades sobre Epilepsia. Accessed on October, 2013. Available from: http://www.epilepsia.pt/lpce/geberalidades-sobre-epilepsia

[Ludwick, 08] Ludwick, D.A. and J. Doucette, Adopting electronic medical records in primary care: Lessons learned from health information systems implementation experience in seven countries International journal of medical informatics, 2008. 78: p. 22-31.

[Musen, 14] Musen, Mark A., Blackford Middleton, and Robert A. Greenes. Clinical decision-support systems. Biomedical informatics. Ed. Shortliffe, E. and Cimino, J. Springer-Verlag London, 2014. p. 643-674.

[Osheroff, 04] Osheroff, J., Pifer, E., Sittig, D., and Jenders, R. Clinical decision support implementers' workbook. Chicago: HIMSS, 2004.

[Romano, 11] Romano, M. J. and Stafford, R. S. Electronic health records and clinical decision support systems: impact on national ambulatory care quality. Arch Intern Med, 2011. 171(10): p. 897-903.

[Roque, 11] Roque, F.S., et al., Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. PLoS Comput Biol, 2011.

[Ruch, 07] Ruch, P. and J. Gobeill, From clinical narratives to ICD codes: automatic text categorization for medico-economic encoding. AMIA - Annual Symposium Proceedings, 2007.

[ICD9Data.com, 13] Software, A. The International Classification of Diseases, 9th Revision, Clinical Modification. May 30th 2013]; Available from:
http://www.icd9data.com/2013/Volume1/320-389/340-349/345/default.htm.

[Tsumoto, 11] Tsumoto, S. and S. Hirano, Clustering-based Analysis in Hospital Information Systems. International Conference on Granular Computing, 2011: p. 669-674.

[Tan, 06] Tan, P.-N., M. Steinbach, and V. Kumar, Introduction to Data Mining. Vol. 1. 2006: Pearson Education.

[WHO, 13] WHO – World Health Organization. Epilepsy, Fact Sheet Nº 999, October 2012. Available from: http://www.who.int/mediacentre/factsheets/fs999/en/

[Witten, 03] Witten, I.H., Text mining. In Practical handbook of Internet computing, 2003. Available from: http://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf

[Tseng, 07] Tseng, Y.H. and C.J. Lin, Text mining techniques for patent analysis. Information Processing and Management, 2006. 43(5): p.1216-1247.

[Wongthongtham, 09] Wongthongtham, P. and E. Chang, Development of a Software Engineering Ontology for Multi-site Software Development. IEEE Transactions on knowledge and Data Engineering, 2009.