Computer Methods and Programs in Biomedicine 165 (2018) 139–149

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia

Leonard Barreto Moreira[a,b,*], Anderson Amendoeira Namen[b,c]

[a] Postgraduate Program in Cognition and Language, North Fluminense State University – UENF, Av. Alberto Lamego, 2000 - Parque Califórnia - CEP 28013-602, Campos dos Goitacazes, Rio de Janeiro, Brazil.
[b] Computer Modelling Department, State of Rio de Janeiro University, Rua Bonfim, 25 - Vila Amélia - CEP 28625-570 - Nova Friburgo, Rio de Janeiro, Brazil.
[c] Veiga de Almeida University, Rua Ibituruna, 108 - Maracanã - CEP 20271-020, Rio de Janeiro, Brazil.

ARTICLE INFO

Article history:
Received 19 May 2018
Revised 31 July 2018
Accepted 21 August 2018

Keywords:
Data mining
Text mining
Alzheimer's disease
Mild cognitive impairment
Medical diagnosis

ABSTRACT

Background and Objective: Given the phenomenon of aging population, dementias arise as a complex health problem throughout the world. Several methods of machine learning have been applied to the task of predicting dementias. Given its diagnostic complexity, the great challenge lies in distinguishing patients with some type of dementia from healthy people. Particularly in the early stages, the diagnosis positively impacts the quality of life of both the patient and the family. This work presents a hybrid data mining model, involving the mining of texts integrated to the mining of structured data. This model aims to assist specialists in the diagnosis of patients with clinical suspicion of dementia.
Methods: The experiments were conducted from a set of 605 medical records with 19 different attributes about patients with cognitive decline reports. Firstly, a new structured attribute was created from a text mining process. It was the result of clustering the patient's pathological history information stored in an unstructured textual attribute. Classification algorithms (naïve bayes, bayesian belief networks and decision trees) were applied to obtain Alzheimer's disease and mild cognitive impairment predictive models. Ensemble methods (Bagging, Boosting and Random Forests) were used in order to improve the accuracy of the generated models. These methods were applied in two datasets: one containing only the original structured data; the other containing the original structured data with the inclusion of the new attribute resulting from the text mining (hybrid model).
Results: The models' accuracy metrics obtained from the two different datasets were compared. The results evidenced the greater effectiveness of the hybrid model in the diagnostic prediction for the pathologies of interest.
Conclusions: When analysing the different methods of classification and clustering used, the better rates related to the precision and sensitivity of the pathologies under study were obtained with hybrid models with support of ensemble methods.

1. Introduction

Population aging is a worldwide phenomenon that occurs at different rates in different regions of the world. Particularly in Brazil, predictions state that in 2050 21.8% of the population will be composed of people aged 60 or over [1]. Human cognition is naturally altered over time, which generates declines in the information processing speed and in the working memory capacity [2]. In this sense, the elderly population growth has the effect of increasing the proportion of adults with some type of dementia. Predictions state that the number of dementia cases will double every 20 years [3].

The medical diagnosis process is not an easy task, given the diversity of diseases, symptoms, examinations, and the complexity of human physiology. The diagnosis of dementia syndromes is based on an objective evaluation of the cognitive and functional performance, which demands an extensive investigation through a series of consultations, evaluations and exams. The complexity is even greater in the early stages, because of the lack of a specific examination that could determine the type of dementia. [4,5].

In diagnostic prediction activities, data mining (DM) and text mining (TM) have been presented as promising methods [6]. These methods are composed by a set of tools and techniques that are

* Corresponding author at: Postgraduate Program in Cognition and Language, North Fluminense State University – UENF, Av. Alberto Lamego, 2000 - Parque Califórnia - CEP 28013-602, Campos dos Goitacazes, Rio de Janeiro, Brazil.
E-mail addresses: leonardbarreto@pq.uenf.br (L.B. Moreira), anamen@uva.br (A.A. Namen).

https://doi.org/10.1016/j.cmpb.2018.08.016
0169-2607/

Table 1
Quantitative and percentages of patients diagnosed with the various types of dementia contained in the data set.

Pathology	Quantitative	Percentage
Alzheimer's disease	290	48.0%
Mild dognitive dmpairment	97	16.0%
Others	218	36.0%

able to explore sets of data, helping to evidence patterns and assisting in the knowledge discovery [7].

Researches aimed at the development of predictive systems for the diagnosis of dementias are significantly increasing. Works such as [8–14] apply several DM techniques for the modelling of dementias predictive systems. However, one of the major challenges for diagnosis is the identification of cases in the early stages, especially in cases of Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI), given the similarity of patients' complaints observed in this phase of clinical research [4,15].

This work presents a hybrid model of mining, involving text mining integrated to the mining of structured data. This model aims to assist the specialists in the diagnosis process of patients with clinical suspicion of early dementia – AD and MCI. It should be noted that despite the possibility of evolution from TCL to other pathologies, in the present study MCI and AD are distinct entities. That is, the MCI patients considered in the study did not progress to AD. The results of the proposed hybrid model are compared to the model containing only the structured data. The goal is to verify the hypothesis that the inclusion into the structured data set, of data resulting from the mining of medical texts, improves the predictive power of the generated models.

2. Materials and methods

The present study used data from the records of 605 patients attended by the Alzheimer and Parkinson Center in the city of Campos dos Goytacazes, located in the State of Rio de Janeiro, Brazil. The distribution of individuals diagnosed with dementia is found in Table 1.

Patient information consisted of demographic, clinical, and screening tests, totalling 19 different attributes, as presented in Table 2. Regarding the neuropsychological instruments of analysis, the scales and cut-off points indicated in the medical literature were adopted, duly based on recommendations established and validated by national and international research institutes [16–25]. The table shows that one of the attributes (18 - PPH) was presented in free, unstructured text format.

The data used for the mining process were limited to the information related to the first attendance (sorting), thus excluding temporal information. Patients with suspected diagnosis of dementia (in the initial stage) whom had a diagnosis established, according to the evolution of the disease, were considered for the data mining process. Those diagnoses were established by the unit physicians after extensive analysis in the medical records. Patient identification data were disregarded at this work, allowing total anonymity and privacy.

2.1. Process overview

Supervised and unsupervised learning methods were used to classify patients with suspected diagnosis of AD and MCI. The process was composed of four main steps: i) Text mining, applying Natural Language Processing (NLP) and Information Retrieval (IR) techniques aiming at textual uniformity, with subsequent application of clustering algorithms (K-means and X-means), in order to create a new attribute, called Cluster_PPH, that grouped the medical reports of the patient's previous pathological history (PPH), which were stored in an unstructured textual attribute; ii) Creation of two data sets: the first containing the attributes presented in Table 2, with the exception of the unstructured textual attribute PPH; the other, also incorporating a new structured attribute (Cluster_PPH), produced in step i, from the unstructured textual attribute PPH. In addition, applying SMOTE (Synthetic Minority Oversampling Technique) to deal with class imbalance in cases of MCI. iii) Development of classification models based on Naïve Bayes (NB), Bayesian Belief Networks (BBN) and Decision Trees (DT) to obtain predictive models, considering the two datasets produced in step ii. Aiming to improve the accuracy of the generated models, we applied ensemble methods (Bagging, Boosting e Random Forests); iv) Comparative analysis of the application of the predictive models on the two sets of data. Metrics such as accuracy, false positive rate (FPR), false negative rate (FNR), sensitivity and area under the ROC curve (AUC) were evaluated. Weka software, a free tool made up of a collection of data mining algorithms and a series of functionalities that aid in the pre-processing step of the data [27], was used to support the various stages of the research. Fig. 1 illustrates the steps performed during the work.

2.2. Mining of textual data using clustering techniques

The adequacy of textual data for clustering purposes requires the use of tasks related to areas such as NLP and IR. These tasks are performed through long and complex iterations, usually supported by external sources of knowledge, in addition to the assistance of domain experts. Recent studies [28,29] investigate the impact on the efficacy of DM algorithms when is included the concomitant use of approaches based on linguistic and mathematical analysis (data-driven approaches) and those based on the domain expert's judgment (knowledge-driven approaches).

2.2.1. Pre-processing tasks

Several tasks for PPH attribute texts pre-processing were performed, aiming to adapt them to later application of clustering techniques. In order to perform these tasks, NLP techniques such as case folding, stopwords removal, stemming, domain ontologies (thesaurus) and tokenization were used, with the application of algorithms available in the Weka tool and algorithms developed by the authors themselves. Firstly, the conversion of the texts' characters to lowercase letters and the removal of accents were carried out from an algorithm implemented by [30].

A typical characteristic in texts is the existence of very common words that have little value for the mining process. These words are represented, in general, by articles, prepositions, conjunctions, or by terms with great frequency throughout the collection of documents [31,32]. The set of words considered irrelevant in this work (called Stopword Lists) was obtained from the Snowball project for the Portuguese language (details in http://snowball.tartarus.org/algorithms/portuguese/stop.txt).

In Portuguese, as in several languages, words occur in the text in more than one form that contain similar meanings – e.g. book (*livro*), books (*livros*), democracy (*democracia*), democratic (*democrático*). In many situations, reducing word's inflectional and derivational forms to a common basic form, by performing morphological analysis in texts, is very effective. This process usually is referred to as stemming [33]. In our work, this process was performed with the support of the open source implementation of the Porter algorithm [34] adapted to Portuguese, called PTstemmer 5 [35].

The incorporation of the knowledge of the theme to the models was carried out with the support of a standard vocabulary and

Table 2
List of attributes used in the data mining process.

Attribute		Status	Description
1	smoker	0: false 1: true 2: uninformed	Indicates whether patient is (was) smoker or not.
2	alcoholic	0: false 1: true 2: uninformed	Indicates whether patient is (was) alcoholic or not.
3	marital_status	0: uninformed 1: single 2: married 3: widower 4: divorced	Patient marital status
4	age	0: < 65 years old 1: > = 65 years old	Patient age
5	gender	1: male 2: female	Patient gender
6	education	0: illiterate 1: 1 to 4 years of schooling 2: 5 to 8 years of schooling 3: 9 or more years of schooling	School years
7	mmse	0: (0–17] 1: [18-26] 2: [27-30]	Mini–mental state examination (MMSE) result (see [17])
8	adl	0: great dependence 1: moderate dependence 2: independence	Activities of Daily Living (ADL) result (see [18])
9	iadl	1: independence 2: partial dependence 3: total dependence	Instrumental Activities of Daily Living (IADL) result (see [19])
10	gds	0: normal 1: depression 2: severe depression	Evaluation result of the Geriatric Depression Scale (see [20])
11	camcog	A: high M: moderate B: low	Result of cognitive impairment index in the CAMCOG test (see [26])
12	cdr	0: healthy 0.5: questionable 1: low 2: moderate 3: significant	Test result for Clinical Dementia Rating (see [22])
13	verbal_fluency	0: < = 10 words 1: > 10 words	Number of words spoken in 1 minute (see [23, 24])
14	diabetes	0: false 1: true	Indicates whether the patient has diabetes
15	ah	0: false 1: true	Indicates if the patient has arterial hypertension
16	cerad	0: score < 13 1: score > ´ = 13	Score obtained in the CERAD test (see [16])
17	clock	0: Neither numbers nor hands are remotely correctly placed 1: Marked visuospatial errors 2: Moderate visuospatial errors 3: Clear-cut errors in time given 4: Mild visuospatial errors 5: Numbers and hands are correctly placed.	Clock drawing test result (see [25])
18	PPH	Unstructured attribute, in text format.	Text, freely filled out by the physician, containing information about the patient's previous pathological history (PPH).
19	diagnosis	A: AD diagnosis M: MCI diagnosis O: other cases	Indicates the patient type of dementia.

hierarchical structures related to pre-established concepts, denominated thesaurus. One of the main benefits of using the thesaurus in document classification systems is the reduction of the terms used for clustering, since words with similar semantic content can be expressed in a unique form. In document clustering this is an important aspect, since polysemic and synonymous words are both relatively prevalent and fundamentally important for the formation of distinct groups [36].

The present work used as a base a structured vocabulary for indexing and retrieval of information related to Health Sciences, called DeCS - Descriptors in Health Sciences [37]. Given the scope of the research, we considered the DeCS existing synonyms related to the dementia syndromes.

Another applied process, called labelling, was characterized by the inclusion of tags in a set of words related to symptoms, comorbidities and risk factors. This procedure aimed to differentiate situations in which the patient was or was not affected by some pathology and/or symptom. Instead of considering the terms individually, we chose a combination of contiguous words (tokens), a strategy that can be useful in identifying certain expressions that occur in the text [38]. With the use of the NgramTokenizer algorithm, available in the Weka tool, the terms with up to three tokens were considered for the mining process.

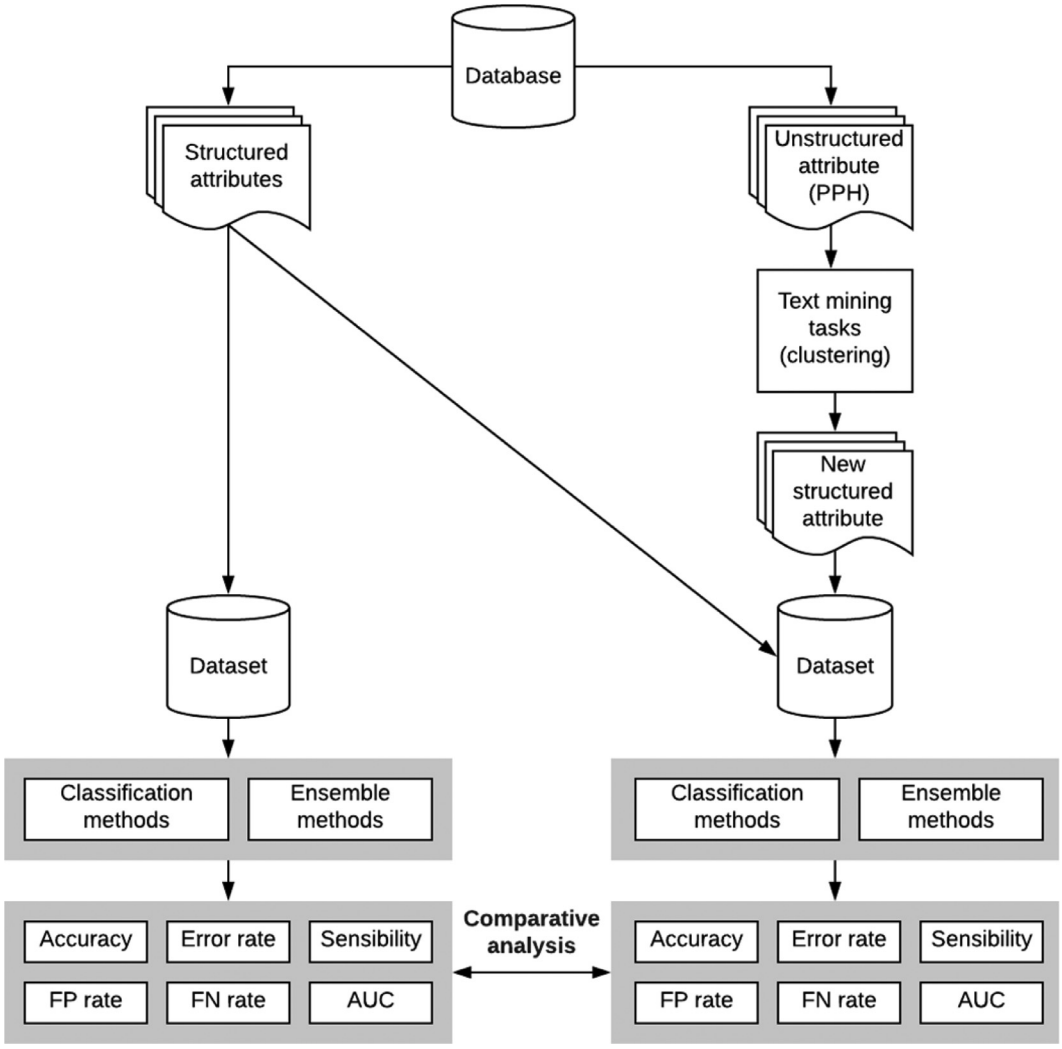

Fig. 1. Data mining and text mining applied to compare predictive models.

Table 3
Set of terms and respective quantitative (between parenthesis), obtained after NLP stage.

	Terms
Most frequent terms	esquec (770); neg_alzheim (336); hipert (334); neg_alzheim famil (292); alzheim (212); neg_diabet (211); apresent esquec (209); onde guard (206); lembr (161); guard objet (148).
Less frequent Terms	acontec recent (21); esta convers (24); alzheim hipert (24); diabet cardiopat (25); ja esquec (25); apresent inson (26); coloc cois (26); dor (26); pacient esquec (26); acompanh relat (27).

After performing the pre-processing steps described above, the number of terms in the texts was reduced from 2258 to 239. Based on this quantitative, a new subset of terms was obtained from the evaluation of two specialist physicians, which chose the terms that they considered relevant to the diagnostic process. As a result, the number was reduced to 106 attributes. Finally, aiming at improving the previously chosen terms, we tried to analyse the list of words in order to agglutinate and/or discard semantically similar or non-discriminatory terms, thus giving a final list containing 50 different terms. Table 3 illustrates the most important features drawn from the medical reports. It presents the 10 most and less frequent terms; this information is relevant because is directly related to the weighting schemes, such as TF-IDF. The complete list containing the 50 terms and their respective quantitative can be found in [30]. Recalling that the terms were extracted from the Portuguese language.

Considered as fundamental problems in TM, the quantification of the frequencies of the terms within the texts, as well as the estimation of these terms relevance were treated, respectively, through the bag-of-words approach [39,40] and the tf-idf weighting scheme [41]. More details can be found in [30].

2.3. Text clustering

Once the pre-processing task was completed, the next step was clustering. Document clustering is a data mining technique that consists of automatically grouping, through unsupervised learning, a set of documents into generally disjoint subsets [40].

The clustering methods are defined by a specific algorithm that partitions the data set, based on some metric of distance or similarity between the objects, aiming to produce clusters, internally homogeneous, and heterogeneous with each other. In TM, one of the most important issues in cluster analysis is related to finding

a function that quantifies the similarity between documents [42]. Traditionally, documents are grouped based on how similar, or dissimilar, are to each other. Given the speed and simplicity of calculating the distance between two documents, researchers often use the Squared Euclidean Distance [Eq. (1)] in TM, considering the normalized data instance vectors. This normalization makes the Squared Euclidean Distance between two instances proportional to the cosine distance [43,44], according to equation:

$$d\left(\vec{v_a}, \vec{v_b}\right) = \sum_{t=1}^{m} \left(w_{t,a} - w_{t,b}\right)^2 \quad (1)$$

where $\vec{v_i}$ denotes the m-dimensional vector on the set of terms $T=\{t_1,.., t_m\}$ representing a given document d_i and $w_{t,d}$ represents the weighting scheme (w) of the terms (t) in the documents d (a and b, in the above equation).

Works about document clustering mostly use partitioned or hierarchical algorithms to organize the documents of a collection into representative groups. Of these, partitional methods are characterized by organizing objects of a set into several unique clusters [7]. Considered one of the most popular methods of partitioning, K-means is one of the most simple and efficient algorithms for unsupervised learning. Among the applications in the area of medicine, we can found works related to the categorization of pathological entities and/or their subtypes, through evaluation of variables related to clinical tests, lifestyle and dietary patterns [45–47].

K-means algorithm is essentially characterized by defining a prototype in terms of centroid, usually applied to objects in an m-dimensional space. The clustering problem can be formally established as follows [7]:

Given a set of data set X having n objects and the number k of clusters, we want to organize the objects in k partitions $C_1,..., C_k$, such that $k \leq n$, $C_i \subset X \wedge C_i \cap C_j = \emptyset$, for $1 \leq i, j \leq k$. Each partition C is represented by a central point (centroid) C_i, where each object x is associated to the nearest centroid by a distance measure, $dist(x, C_i)$. Once the assignment is made, the centroids are updated in each cluster and the whole process is repeated until it reaches a stop criterion.

The clustering models were obtained with the implementation of K-means [48] available in Weka, as well as one of its most prominent variants, called X-means [49]. The evaluation of the produced clusters was carried out with the support of the coefficients related to the cohesion (measured by the within cluster sum of squares - SSE), the separation (through the measure called Sum of Squares Between - SSB), and the Silhouette Coefficient [50].

As a result of the text mining, 4 distinct clusters were generated in our work (details of the process can be found in [30]). Based on the clustering, a new attribute called Cluster_PPH was created. It was composed of 4 distinct values (*cluster1, cluster2, cluster3* and *cluster4*). The attribute value indicated the respective cluster to which the record containing the text data in the PPH attribute referred. As mentioned, this new attribute was incorporated into the original data set, aiming to evaluate if the predictive model's effectiveness would be improved.

2.4. Data mining models

Classification refers to the problem of categorizing observations into classes. Predictive modelling uses data samples for which class is known to generate a model that can classify new observations.

Briefly, the model obtained from n training tuples $X=(x_1,..., x_n)$, is used to label unknown records according to m distinct classes $C=(c_1,..., c_m)$. In this work the following classification methods were used: Naïve Bayes, Bayesian Belief Networks and Decision Trees. Those methods were chosen because such techniques present a good performance in diagnostic prediction tasks in medicine (see works [8–12], mentioned previously). Ensemble methods were also applied to improve the precision of the models.

2.4.1. Naïve Bayes

We want to determine the probability that a tuple X belongs to a class C, considering known data of X. In Bayesian terms, X is considered evidence and is generally described by a set of n attributes of the type $X=(x_1,..., x_n)$. H is considered as the hypothesis where the data of the tuple X belong to a specific class of C. Therefore, we should determine the probability $P(H|X)$. One of the main characteristics of this classifier is the assumption that attributes are conditionally independent, given the label of class C [51]. The assumption of conditional independence implies that:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i) \quad (2)$$

To predict the class label of a tuple X, $P(X|C_i)P(C_i)$ is evaluated for each class C_i. The classifier predicts if X belongs to C_i, considering m possible classes, if and only if:

$$P(X|C_i)P(C_i) > P\left(X \vee C_j\right)P\left(C_j\right), \forall 1 \leq j \leq m, j \neq i \quad (3)$$

In other words, the predicted class label is one in which the class $C_i$for $P(X|C_i)P(C_i)$has the maximum value.

2.4.2. Bayesian belief networks

BBNs are characterized by the possibility of representing dependencies between a subset of variables [52]. They are usually represented by a Directed Acyclic Graph (DAG) in which the nodes represent the variables ($Y_1,..., Y_n$)and the edges the dependency relationships between a set of variables. Each model variable has a conditional probability table, each table specifying the conditional distribution in relation to its immediate parent nodes, given by Eq. (4):

$$P(x_i, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \vee Parents(Y_i)) \quad (4)$$

where $P(x_i,..., x_n)$ is the probability of a particular combination of values of X, $X=(x_1,..., x_n)$ is a data tuple described by the variables or attributes $Y_1,..., Y_n$, respectively, and the values for $P(x_i \vee Parents(Y_i))$ correspond to the entries in the DAG for Y_i [51].

The modelling process of a Bayesian network involves a series of algorithms, highly dependent on the type of data and the way in which causal relations will be represented. One should evaluate the type of data (discrete or continuous) and the existence (or not) of hidden variables. The BBN structure can be built by domain experts or obtained automatically from the training data. Hybrid approaches based on constraint-based and score-based methods are used to estimate the parameters for a given BBN structure. With respect to the proposed BBN in this work, only discrete values were considered and the structure was inferred from the training data with support of the ICS (Inductive Causation Search) algorithm, available in Weka.

2.4.3. Decision trees

A decision tree (DT) is a structure that divides successively a set of records into smaller and smaller sets, adopting the divide-to-conquer strategy. In a DT, the root node and the inner nodes contain a test condition about an attribute, the branches represent the values of this attribute, and the terminal nodes (leaves) receive the classes. Thus, from the root node, the classification of a new tuple is performed by applying the test condition to each node, which will direct to the appropriate branch to the leaf node.

Two factors are crucial for the construction of an AD: (1) the way in which the attributes are selected and divided for tree generation, and (2) the mechanism used to prune the tree [51].

Table 4
Quantitative and percentages of test and training subsets for AD and MCI pathologies before and after the application of SMOTE.

Pathology	Original Data set						Data set after application of SMOTE (only MCI)		
	Total records for each class (test set)			Total records for each class (training set)			Total records for each class (training set)		
	diagnosis = 0	diagnosis = 1	Total	diagnosis = 0	diagnosis = 1	Total	diagnosis = 0	diagnosis = 1	Total
AD	79 (52.0%)	73 (48.0%)	152	236 (52.0%)	217 (48.0%)	453	–	–	–
MCI	127 (83.6%)	25 (16.4%)	152	381 (84.1%)	72 (15.9%)	453	381 (50.0%)	381 (50.0%)	762

Note: value 1 indicates occurrence of pathology and 0 indicates non-occurrence

Among the heuristics to select the division criterion that best separates a data partition, the information gain (algorithm ID3 [53]) and the gain ratio (algorithm C4.5 [54]) can be cited, the latter considered an extension of the information gain technique. The possible scenarios of division and creation of the branches (or leaf nodes) are based on the types of attributes, being characterized by having multiple or binary splits.

Two approaches are used to prune DTs in order to address problems of data overfitting: pre-pruning or post-pruning. Pre-pruning involves the decision to stop generating ramifications during the construction of the decision tree, according to statistical measures. In the post-pruning approach, the tree optimization process is performed after its complete construction, using subtree raising and subtree replacement methods. The basic idea of the subtree replacement strategy is to select some subtree and replace it with a leaf, performing tests from the leaves to the root node. When subtree raising is implemented, the method selects a subtree and replaces it with a child, based on a stipulated threshold for the node/leaf error rate.

In this work, we used the algorithm called j48, an implementation of the algorithm C4.5 available in the Weka tool. The selection of attributes for the nodes followed the Gain Ratio criterion, whose highest value attribute is selected as the division attribute. The binary approach was used to divide the nodes, generating a more representative and easy-to-understand model for domain specialists. In order to create a less complex tree, we decided to perform pruning after creating the tree structure, using the subtree raising method.

2.4.4. Ensemble methods

Ensemble methods are techniques that aggregate predictions from multiple classifiers with the aim of improving the accuracy of prediction models [55]. The adoption of these methods has being grown in the medical field (see [8,9,56,57]). In general, the combination of k classification models, denoted by $M_i, i = 1, ..k$, helps to compile a composite classification model. A given data set, D, is used to create k training sets, $D_1, D_2, ..., D_k$, where $D_i (1 \leq i \leq k-1)$, is used to generate classifier M_i. The classification of a new tuple is obtained by some type of combination (usually by weighted or unweighted voting) from each base classifier [51]. Among these methods are the techniques of Bagging [58], Boosting [59] and Random Forests [60], all of which are used in the present work.

3. Results

For the conduction of the experiments, the original data set (605 records) was replicated aiming to evaluate each of the two pathologies, AD and MCI. In each data set, the value of the diagnosis attribute was changed: value 1 was set in cases of pathology diagnosis, 0 in other cases. Further, both data sets were partitioned into two mutually exclusive subsets called the training set and test set in the ratio ¾ e ¼, respectively. As observed in Table 4, given the small number of positive cases for MCI, the balancing method called SMOTE [61] was used to modify the training set referring to this pathology, adding synthetic records of minority class to make the training set evenly distributed.

Table 5
Evaluation metrics.

Measure	Formula	Domain	Best score
Accuracy	$(TP+TN)/(P+N)$	[0,1]	1
Sensitivity	TP/P	[0,1]	1
FNR	1-Sensitivity	[0,1]	0
FPR	$FP/(FP+TN)$	[0,1]	0
AUC	–	[0,1]	1

3.1. Evaluation metrics

Aiming to quantify and compare the performance of the classification methods, the following metrics were used: accuracy, sensitivity, false negative rate (FNR), false positive rate (FPR) and area under the ROC curve (AUC). Table 5 presents the metrics that were used to evaluate the quality of the developed models. The counting of positive records correctly classified by the model (TP - true positives), negative records correctly classified by the model (TN - true negatives), negative records erroneously labelled as positive (FP - false positives), and positive records incorrectly classified as negative (FN - false negatives) serve as the basis for calculating the above measures. It should be noted that for the present study, a registry classified as positive indicates a positive diagnosis of the pathology (negative classification indicating the inverse situation).

3.2. Models performance evaluation

As mentioned in Section 2, the ensemble methods Bagging, AdaBoost and Random Forests were applied in order to improve the accuracy of the models. The parameters used in the models were estimated based on the optimizer described in [62], available in the Weka tool. The strategy for the definition of such values was based on two steps: (1) definition of the best parameters for the base classifiers; (2) once set the parameters obtained in the first step, choosing the best set of values for the ensemble methods. All the parameters values used are presented in [30].

An important issue in machine learning relates to how the errors produced by the models are calculated. In some cases, the probability analysis of each forecast can be considered a complementary feature to the error-based metrics [63]. In most techniques, this calculation consists of the difference between the actual and predicted result, from the application of equations known as loss functions. Depending on the machine learning task different loss functions can be used. The loss function used in the Weka tool for binary classifiers is the log loss. As with error-based metrics, the use of method [62] improved the model reliability by incorporating such probabilistic information. After the training phase of the model, a threshold, defined as 0.5, was established aiming at maximizing accuracy.

For each pathology, we evaluated the models generated from two different data sets: one considering the original set of attributes, excepting the textual attribute PPH, which referred to a

Table 6
Results of the classification techniques for the dataset with and without the PPH attribute, for AD diagnosis. Best results highlighted in bold.

Clustering method	Ensemble method	Classifier	Accuracy	Sensitivity	FNR	FPR	AUC
None – PPH attribute not considered	–	j48	0.724	0.781	0.219	0.329	0.722
	–	NB	0.658	0.767	0.233	0.443	0.716
	–	BBN	0.645	0.795	0.205	0.494	0.679
	AdaBoost	j48	0.730	0.712	0.288	0.253	0.818
		NB	0.618	0.644	0.356	0.405	0.697
		BBN	0.638	0.781	0.219	0.494	0.631
	Bagging	j48	0.710	0.740	0.260	0.316	0.818
		NB	0.664	0.753	0.247	0.418	0.723
		BBN	0.664	0.753	0.247	0.418	0.685
	Random Forests	–	0.724	0.753	0.247	0.304	0.800
KMeans ($k=4$)	–	j48	0.737	0.822	0.178	0.342	0.789
	Adaboost	j48	0.730	0.726	0.274	0.266	0.806
X-means	–	j48	0.737	0.822	0.178	0.342	0.789
	Adaboost	j48	**0.796**	**0.836**	**0.164**	**0.241**	**0.849**

Table 7
Results of the classification techniques for the data set using SMOTE for class balance, with and without considering the PPH attribute, for MCI diagnosis. Best results highlighted in bold.

Clustering method	Ensemble method	Classifier	Accuracy	Sensitivity	FNR	FPR	AUC
None – PPH attribute not considered	–	j48	0.862	0.320	0.680	0.031	0.813
	–	NB	0.829	0.760	0.240	0.157	0.862
	–	BBN	0.809	0.520	0.480	0.134	0.831
	AdaBoost	j48	**0.901**	0.512	0.488	0.024	0.869
		NB	0.877	0.400	0.600	0.031	0.804
		BBN	0.855	0.400	0.600	0.055	0.817
	Bagging	j48	0.842	0.120	0.880	**0.016**	0.858
		NB	0.835	0.760	0.240	0.150	0.863
		BBN	0.862	0.760	0.240	0.150	0.863
	Random Forests	–	0.849	0.280	0.720	0.039	0.866
K-means ($k=4$)	–	NB	0.842	**0.800**	**0.200**	0.150	0.872
	Bagging	NB	0.849	**0.800**	**0.200**	0.142	**0.874**
X-means	–	NB	0.842	**0.800**	**0.200**	0.150	**0.874**
	Bagging	NB	0.849	**0.800**	**0.200**	0.142	**0.874**

text prepared by the physician, in free format, with comments about the patient; the other (hybrid model), considering the insertion of the Cluster_PPH attribute, which contained the result of the clustering done, through the text mining over the PPH attribute. Remembering that the Cluster_PPH attribute could have 4 distinct values, indicating the respective cluster to which the record containing the text data in the PPH attribute referred.

The results obtained for the interest classes were tabulated (Tables 6 and 7) aiming to facilitate the comparison between the values obtained from the classification and clustering methods under the two different approaches. Due to the large number of hybrid models generated (data set with the Cluster_PPH attribute) from the use of two different clustering methods (K-means and X-means), only the best results of these models' simulations will be presented.

3.2.1. Evaluation of predictive models for AD

According to the values summarized in Table 6 for the diagnosis of AD, without considering the inclusion of the structured attribute derived from the text mining, j48 presented not only the best rate of correctly classified instances (accuracy), as well as a better relationship between sensitivity and FPR, observed at the AUC value. However, the classifier BBN presented slightly better rates for the sensitivity and FNR metrics. Analysing the results of the ensemble methods in the same table, from the point of view of sensitivity and FNR, no improvement was obtained. Nevertheless, improvements in accuracy, FPR and AUC were obtained through the application of AdaBoost associated to the classifier j48. In relation to the AUC metric, improvement is also observed for this classifier associated with the Bagging ensemble method. The rate of individuals misdiagnosed as not AD, that is, the FNR, was not minimized with the application of ensemble methods.

The lines of Table 6 that indicate the clustering methods shows the metrics related to the accuracy of the AD diagnosis models, considering the inclusion of the Cluster_PPH attribute, resulting from TM. It is noticed that the decision tree technique j48 and the ensemble methods Random Forests and AdaBoost, the latter having j48 as the base classifier, presented the best rates of correctly classified instances (accuracy) in the experiments performed with all clustering methods. Specifically associated to X-means, all the metrics considered in the present work for the mentioned pathology had improvements in comparison with the previously obtained results without Cluster_PPH, that is, the insertion of the new attribute, result of the process of TM, implied in the improvement in the precision of the results.

3.2.2. Evaluation of predictive models for MCI

As can be observed in Table 7, with respect to the diagnosis of MCI without considering the inclusion of the Cluster_PPH attribute in the dataset, the J48 classifier presented the best FPR rates and accuracy for the MCI diagnostic hypothesis. Still in this group of classifiers, a substantial improvement in FNR was observed through the Naïve Bayes method. By analysing the ensemble methods, improvements in rates such as accuracy and AUC were observed in general, with AdaBoost, using j48, presenting the highest rates of the previously mentioned metrics. For Bagging with Naïve Bayes and BBN, it is observed that the FNR values were the same as those of the classifier Naïve Bayes. However, there is a substantial improvement in the other rates (FPR, AUC and accuracy).

Table 8
Decision rules for the diagnosis of AD obtained through the hybrid model.

Rule	Rule Description
R1	if (mmse = 1) and (Cluster_PPH! = cluster4) and (diabetes = 1) and (camcog! = B) then diagnosis = 1 (26/10)
R2	if (mmse = 1) and (Cluster_PPH! = cluster4) and (diabetes = 1) and (camcog = B) and (gds = 0) then diagnosis = 1 (45/20)
R3	if (mmse = 0) and (Cluster_PPH! = cluster4) and (verbal_fluency = 0) then diagnosis = 1 (107/45)
R4	if (mmse = 0) and (Cluster_PPH! = cluster4) and (verbal_fluency! = 0) and (education! = 1) then diagnosis = 1 (41/14)

Table 7 also presents a selection of the best results for predicting MCI, in terms of precision, from the application of the different methods, considering the inclusion of the Cluster_PPH attribute in the data set. The naïve bayes technique, as a base classifier in conjunction with the Bagging method, presented the best metrics in experiments with all clustering methods. Also, the X-means and K-means ($k = 4$) clustering methods appear as the best models when considering the Cluster_PPH attribute for the mining process. Despite of lower accuracy values, the other metrics, excepting FPR, had improvements evidencing an enhancement of the results in this approach.

3.3. Patterns related to the clusters

As observed in the results, the models that considered the inclusion of the Cluster_HPP attribute, for both AD and MCI, presented results with greater precision. In this sense, it is believed that the information presented by physicians in free and unstructured format (HPP attribute) contained significant data related to patients. However, since this information was not structured, it would be discarded and not used in the predictive model. By including the process of grouping and structuring these data, it is believed that rich information has been added to the resulting models, which generated gains in terms of precision.

Aiming to interpret each of the four clusters obtained, a supervised learning process was applied over 51 attributes, one of them being the Cluster_HPP attribute, considered the target attribute in the classification model. The remaining 50 attributes corresponded to the different terms used in the clustering process. The decision tree algorithm C4.5, whose implementation in the Weka is denominated j48, was applied. The percentage of records correctly classified by the model was 99.34% (see [30]). It was possible, therefore, to list the rules related to the classificatory model obtained.

The rules presented in [30] are not listed here, as they show terms in Portuguese language and would be difficult to understanding. However, we present a summary of the patterns found for each cluster from the rules' analysis. They indicate that individuals in *cluster1* have mood disorders. In addition, they are less prone to hypertension. *Cluster2* patients do not tend to have irritable behaviour. *Cluster3* individuals present mood disorders and diabetes as a systemic disease. Pathologies such as diabetes are present, with less intensity, in *cluster4* reports. However, when associated with hypertension, diabetes occurs with more intensity. It is also considered, for this cluster, a lower incidence of cardiovascular diseases. Finally, *cluster4* patients also have mood changes and significant forgetfulness related to memory.

3.4. Patterns related to diagnoses of clinical suspicion of dementia

Based on Tables 6 and 7 analyses, the hybrid approach (TM + DM), considering the resulting model of the X-means clustering scheme in conjunction with the AdaBoost ensemble method and the classifier j48, obtained the highest accuracy for the AD clinical suspicion diagnosis. The model's patterns related to the AD diagnosis (*diagnosis = 1*) are listed in Table 8 and are summarized in the rules induced by the classifier. It should be noted that two values are presented between parentheses after each rule; the first represents the number of records correctly classified for the class of interest, while the second value expresses the number of records, among the classified, that received the label of the class erroneously.

Table 9
Hybrid model generated by Naïve Bayes technique for MCI pathology.

Attribute		Value	Probability
1	smoker	0	0.960
2	alcoholic	0	0.980
3	marital_status	0	0.520
4	age	1	0.900
5	gender	2	0.750
6	education	1	0.800
7	mmse	1	0.800
8	adl	2	0.980
9	iadl	1	0.980
10	gds	0	0.700
11	camcog	B	0.900
12	cdr	0.5	0.700
13	verbal_fluency	1	0.870
14	diabetes	1	0.610
15	ah	1	0.850
16	cerad	0	0.660
17	clock	2	0.240
18	Cluster_PPH	cluster2	0.620

Among the main patterns observed for the AD diagnosis, the relevance of the screening tests (mmse and verbal_fluency) as well as the clusterings of the medical reports of the PPH stand out.

All rules have the value of Cluster_PPH attribute not equal to cluster4. This means that, for all rules, the increasing of pathologies such as diabetes does not occur because of its association with hypertension. Also, patients do not have necessarily a situation of lower incidence of cardiovascular diseases combined with mood changes and significant forgetfulness related to memory. R1 rule suggests that the pathology affects individuals with median MMSE (*mmse = 1*) and diabetes. Furthermore, these patients have no indicative of cognitive impairment (*camcog = B*). R2 indicates that non-depressive (*gds = 0*) patients with median cognitive impairment (*mmse = 1*), diabetics (*diabetes = 1*) and low cognitive impairment in CAMCOG (*camcog = B*) tend to AD diagnosis. Rules R3 and R4 differ mainly by a lower score in MMSE (*mmse = 0*). R3 also shows little individual's ability to spontaneously produce words of a given category at a given period of time (*verbal_fluency = 0*). Patients with good verbal fluency (*verbal_fluency! = 0*) but with a median level of education (R4) tend to be affected by this pathology.

The MCI diagnosis obtained the highest accuracy from the model represented by the X-means clustering scheme in conjunction with the Bagging ensemble method, whose classifier was the naïve bayes. Because of space limitations, the patterns related to the model, listed in Table 9, present only the highest percentages related to the value of each attribute, for individuals affected by MCI (*diagnosis = 1*).

By analysing Table 9, among the characteristics related to risk factors, age was pointed out as a critical aspect. Among the set

of attributes considered modifiable, hypertension ($ah = 1$) and low level of schooling ($education = 1$) appear as determining factors for the diagnosis. Regarding the screening tests, we note that the individuals affected by MCI are independent for the daily life activities ($adl = 2$), as well as have the ability to manage the living environment inside and outside the home ($iadl = 1$). A low cognitive impairment ($camcog = B$) is highly probable (90%) in this group. The ability to produce words in a given period of time is not affected in these patients, a fact observed by the verbal fluency test ($verbal\_fluency = 1$). Among the medical reports contained in the PPH field, patients diagnosed with MCI do not tend to have irritable behaviour (i.e., $Cluster\_PPH = cluster2$).

4. Discussion

The purpose of the present study was to propose a decision support system for the diagnosis of dementias in patient triage, considering patients with cognitive problems in the initial stages. The mining methods used, as well as the generated models, were presented for the diagnostic prediction of dementias, with AD and MCI being the diseases of interest. The theme is justified by its relevance given the high incidence of dementia among the elderly, especially over 65, being one of the main causes of morbidity in this age group. The increase in the population's life expectancy further enhances this problem.

We believe that the greatest contribution of the work was the proposal of a hybrid approach, involving the application of text mining in conjunction with structured data mining. This approach allowed the generation of greater precision predictive models when compared to the traditional data mining approach.

Considering the original set of attributes and the AD class of interest, we concluded that ensemble methods, in most cases, obtained improvements in the accuracy of the models. The best values obtained for FPR and FNR were achieved by Bayesian classification techniques and decision trees using ensemble methods. In relation to the AUC metric, the J48 algorithm presented the best rates with AdaBoost and Bagging. After the integration of the new structured attribute Cluster_PPH, result of the text mining process, a substantial improvement in the correctness rate was observed and, mainly (given its importance in the medical domain), in the number of false positives. A similar behaviour was observed, but to a lesser extent, when evaluating the FNR and AUC metrics. The best model was obtained with the AdaBoost ensemble method, having as base classifier the J48 algorithm and considering the grouping of the medical narratives (PPH text attribute) from the clustering algorithm X-means.

For the MCI data mining which accessed the original attributes, we concluded that the ensemble methods, in most cases, obtained a slight improvement in the metrics results in comparison to the base classifiers. From the point of view of accuracy, Bayesian techniques presented an increase in their rates with the use of ensemble methods, as well as the J48 algorithm associated with the AdaBoost. The best values obtained for FPR and FNR were achieved by Bayesian classification techniques alone and in conjunction with ensemble methods. With respect to the metric AUC, the J48 algorithm presented the best rate in conjunction with AdaBoost. However, the relationship between the true negative rate (known as specificity) and the true positive rate (i.e. sensitivity), denoted by the AUC metric, did not depict a good model, as suggested by the obtained values. We observed in this case, despite the good rate associated with FPR, an average sensibility of the model. Such a balance between FNR and FPR was achieved from Bayesian networks with the Bagging ensemble method. When considering the Cluster_PPH attribute integrated to the previously mentioned data set, an improvement was observed in FNR as well as in the sensitivity versus specificity relationship denoted by the AUC metric. Naïve Bayes associated to Bagging presented the best performance in both analyses. Finally, the best model was achieved with the Bagging ensemble method, based on the Naïve Bayes algorithm, considering the structuring and respective grouping of the textual attribute PPH with the K-means ($k = 4$) and X-means clustering algorithms.

By analysing the hypothesis of improvement in the hybrid model results, we concluded that the clustering of the texts written in free format by the physicians (PPH) and the subsequent integration to the previously used data, improved accuracy of predictive models in all pathologies. The hybrid model for AD presented better rates related to accuracy and sensitivity compared to the works of [10,65]. Similar behaviour can be observed for MCI when comparing the metrics related to accuracy and sensitivity with the work of [66]. Those studies were conducted with information regarding control patients at different stages of the disease, thus contrasting with the characteristics of the present study, which analysed patients with cognitive complaints in the initial stages.

For the situation related to the diagnosis of a disease, some might argue that it is preferable to have higher values of FPR than of FNR. The first one indicates false alerts, that is, wrong disease diagnoses, while the second indicates diagnoses that should be given (but were not). By analysing this specific medical domain, it could be believed that having a disease diagnosis would be better, even if wrong, since additional tests could be done to confirm it. On the other hand, for the FNR metric a false negative would indicate the possibility of not continuing the procedures with the patient, which could imply in the evolution of the disease without medical support and monitoring. Considering this perspective, a possible new approach could consider different costs for the FPR and FNR metrics, aiming to adjust the threshold to be used in the classification models. Remembering that, in the present work, the threshold was 0.5 which optimizes the costs of FP and FN equally [64].

As a suggestion of future work, other techniques for the treatment of medical texts contained in the PPH attribute could be applied and evaluated, as well as other unsupervised clustering methods, for example K-Medoids. The richness of texts, their dimensionality, and the inherent complexity of human language, together impose a series of challenges for the modelling of textual documents. Thus, other ways of measuring the importance of terms in the collection of documents could also be evaluated, such as the relevance score [67], the correlation coefficient [68] and the information gain [69], as well as the use of other NLP and IR techniques for extraction of characteristics contained in medical texts, such as POS Tagging. We also suggest the use of other classification techniques, such as support vector machines and neural networks, and ensemble methods such as Stacking [70], in order to evaluate the impact on the accuracy of models Finally, we recommend evaluating the use of algorithms based on deep learning approaches, considering the increasing adoption of such methods in biomedical research. It is believed that this kind of investigation could increase the results accuracy.

Conflict of interest statement

The authors declare that they have no conflict of interest.

The authors declare that the research is in conformity with the recommendations of the Brazilian National Research Ethics Council.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2018.08.016.

References

[1] United Nations Population Fund, Ageing in the Twenty-First Century: A Celebration and A Challenge, United Nations Population Fund (UNFPA), New York, 2012.

[2] E.M. Zelinski, S.E. Dalton, S. Hindin, Cognitive changes in healthy older adults, J. Am. Soc. Aging 35 (2011) 13–20.

[3] M. Guerchet, M. Prina, M. Prince, Policy brief for heads of government: the global impact of dementia 2013–2050, Policy Br Heads Gov Glob Impact Dement 2013–2050 Publ by, Alzheimer's Dis Int (ADI), London, 2013 December 2013 1–8.

[4] EVde Freitas, L. Py, Tratado De Geriatria e Gerontologia (Geriatrics and Gerontology Treaty), 4th ed, Guanabara-Koogan, Rio de Janeiro, 2016.

[5] B. Dubois, A. Padovani, P. Scheltens, et al., Timely diagnosis for alzheimer's disease: a literature review on benefits and challenges, J. Alzheimer's Dis. 49 (2016) 617–631, doi:10.3233/JAD-150692.

[6] S.H. El-Sappagh, S. El-Masri, A distributed clinical decision support system architecture, J. King Saud Univ. Comput. Inf. Sci. 26 (2014) 69–78, doi:10.1016/j.jksuci.2013.03.005.

[7] P-N Tan, M. Steinbach, V. Kumar, Introdução Ao Data Mining, 1st ed, Ciência Moderna, Rio de Janeiro, 2009.

[8] S. Joshi, D. Shenoy, et al., Classification of Alzheimer's Disease and Parkinson's Disease by Using Machine Learning and Neural Network Methods, in: 2010 Second Int Conf Mach Learn Comput, 2010, pp. 218–222, doi:10.1109/ICMLC.2010.45.

[9] P.C. Austin, V. Tu J, J.E. Ho, et al., Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes, J. Clin. Epidemiol. 66 (2013) 398–407, doi:10.1016/j.jclinepi.2012.11.008.

[10] J. Williams, A. Weakley, Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia, in: Expand Boundaries Heal Informatics Using Artif Intell Work Assoc Adv Artif Intell Conf, 2013, pp. 71–76.

[11] F.L. Seixas, B. Zadrozny, J. Laks, et al., A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment, Comput. Biol. Med. 51 (2014) 140–158, doi:10.1016/j.compbiomed.2014.04.010.

[12] F.L. Ferreira, S. Cardoso, D. Silva, et al., Improving prognostic prediction from mild cognitive impairment to Alzheimer's disease using genetic algorithms, in: 11th International Conference on Practical Applications of Computational Biology & Bioinformatics. PACBB 2017, 2017, pp. 180–188.

[13] S.H. Wang, Y. Zhang, Y.J. Li, et al., Single slice based detection for Alzheimer's disease via wavelet entropy and multilayer perceptron trained by biogeography-based optimization, Multimed. Tools Appl. Vol 77 (Issue 9) (2018) 10393–10417, doi:10.1007/s11042-016-4222-4.

[14] S-H Wang, S. Du, Y. Zhang, et al., Alzheimer's disease detection by Pseudo Zernike moment and linear regression classification, CNS Neurol. Disord. Drug Targets 16 (2017) 11–15, doi:10.2174/1871527315666161111123024.

[15] M.S. Albert, S.T. DeKosky, D. Dickson, et al., The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, Focus (Madison) 11 (2013) 96–106.

[16] P.H.F. Bertolucci, I.H. Okamoto, S.M.D. Brucki, et al., Applicability of the CERAD neuropsychological battery to Brazilian elderly, Arq. Neuropsiquiatr. 59 (2001) 532–536.

[17] S.M.D. Brucki, R. Nitrini, P. Caramelli, et al., Sugestões para o uso do mini-exame do estado mental no Brasil (Suggestions for the use of mini-mental state in Brazil), Arq. Neuropsiquiatr. 61 (2003) 777–781, doi:10.1590/S0004-282X2003000500014.

[18] Y.A. Duarte, O. de, C.L. de Andrade, M.L. Lebrão, O Índex de Katz na avaliação da funcionalidade dos idosos (The Katz Index in assessing the functionality of the elderly), Rev. da Esc. Enferm. 41 (2007) 317–325, doi:10.1590/S0080-62342007000200021.

[19] M.P. Lawton, M. Moss, M. Fulcomer, M.H. Kleban, A research and service oriented multilevel assessment instrument, J. Gerontol. 37 (1982) 91–99, doi:10.1093/geronj/37.1.91.

[20] J.I. Sheikh, J.A. Yesavage, Geriatric Depression Scale (GDS) recent evidence and development of a shorter version, Clin. Gerontol. 5 (1986) 165–173, doi:10.1300/J018v05n01_09.

[21] L. Ericeira-valente, C. Tiel, C. Eduardo, et al., CAMCOG – Valores das subescalas em idosos normais com níveis diferentes de escolaridade – Aspectos preliminares (CAMCOG - Values of subscales in normal elderly with different levels of schooling - preliminary aspects, Rev. Bras. Neurol. 49 (2013) 32–36.

[22] M.B.M. Macedo Montaño, S. Andreoni, L.R. Ramos, Clinical dementia rating independently predicted conversion to dementia in a cohort of urban elderly in Brazil, Int. Psychogeriatrics 25 (2013) 245–251, doi:10.1017/S1041610212001615.

[23] R. Nitrini, P. Caramelli, C.M.C. Bottino, et al., Diagnóstico de doença de Alzheimer no Brasil: avaliação cognitiva e funcional (Diagnosis of Alzheimer's disease in Brazil: cognitive and functional evaluation), Arq. Neuropsiquiatr. 63 (2005) 713–719, doi:10.1590/S0004-282X2005000400034.

[24] L. Márcia, C. Claudia, S. Claudia, et al., Doença de Alzheimer: Avaliação cognitiva, comportamental e funcional (Alzheimer's disease: cognitive, behavioral, and functional assessment), Dement. Neuropsychol. 5 (2011) 21–33.

[25] K.I. Shulman, R. Shedletsky, I.L. Silver, The challenge of time: clock-drawing and cognitive function in the elderly, Int. J. Geriatr. Psychiatry 1 (1986) 135–140, doi:10.1002/gps.930010209.

[26] E.M.P. Paradela, C.D.S. Lopes, R.A. Lourenço, Reliability of the Brazilian version of the Cambridge cognitive examination revised CAMCOG-R, Arq. Neuropsiquiatr. 67 (2009) 439–444, doi:10.1590/S0004-282X2009000300013.

[27] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed, Morgan Kaufmann, San Francisco, CA, USA, 2011.

[28] J. Sun, J. Hu, D. Luo, et al., Combining knowledge and data driven insights for identifying risk factors using electronic health records, AMIA Annu. Symp. Proc. 2012 (2012) 901–910.

[29] T.H. Cheng, C.P. Wei, V.S. Tseng, Feature selection for medical data mining: comparisons of expert judgment and automatic approaches, in: Proceedings - IEEE Symposium on Computer-Based Medical Systems, 2006, pp. 165–170.

[30] Moreira LB (2016) Um modelo híbrido de mineração de dados para suspeita diagnóstica relacionada a síndromes demenciais (An hybrid data mining model for dementia diagnosis). Universidade Estatual do Rio de Janeiro. Available in http://www.bdtd.uerj.br/tde_busca/arquivo.php?codArquivo=10544.

[31] BuccioE Di, Maria G, D. Nunzio, et al., Unfolding Off-the-shelf IR Systems for Reproducibility, in: Proc. SIGIR Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR 2015), 2015.

[32] M. da Silva Conrado, V.A. Laguna Gutiérrez, S.O. Rezende, et al., Evaluation of Normalization Techniques in Text Classification for Portuguese, in: B. Murgante, O. Gervasi, S. Misra, et al. (Eds.), Computational Science and Its Applications – ICCSA 2012: 12th International Conference, Salvador de Bahia, Brazil, June 18-21, 2012, Proceedings, Part III, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 618–630.

[33] S. Weiss, N. Indurkhya, T. Zhang, Fundamentals of Predictive Text Mining, 2nd ed, Springer, New York, NY, USA, 2010.

[34] M.F. Porter, An algorithm for suffix stripping, Progr. Electron. Libr. Inf. Syst. 14 (1980) 130–137, doi:10.1108/eb046814.

[35] Acosta O, Aguilar C, Abdullah N, et al (2015) Natural Language Processing and Cognitive Science: Proceedings 2014. 314.

[36] S. Fodeh, B. Punch, P-N Tan, On ontology-driven document clustering using core semantic features, Knowl. Inf. Syst. 28 (2011) 395–421, doi:10.1007/s10115-010-0370-4.

[37] Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BIREME), Organização Pan-Americana da Saúde (OPAS), Organização Mundial da Saúde (OMS), Health Sciences Descriptors: DeCS, 2015 Accessed 23 Nov 2015. http://decs.bvsalud.org/I/homepagei.htm.

[38] Aggarwal CC, Zhai C (2012) Mining text data, 1st ed. Springer-Verlag New York, New York, NY, USA.

[39] F. Provost, T. Fawcett, Data Science for Business: What You Need to Know About Data Mining and Data-analytic Thinking, 1st ed, O'Reilly Media, Inc, 2013.

[40] R. Baeza-Yates, B. Ribeiro-Neto, Recuperação De Informação: Conceitos e Tecnologia Das Máquinas de Busca (Information Retrieval: Search Engine Concepts and Technology, 2a, Bookman Editora, Porto Alegre, 2013.

[41] G. Salton, C.S. Yang, On the specification of term values in automatic indexing, J. Doc. 29 (1973) 351–372, doi:10.1108/eb026562.

[42] S. Owen, R. Anil, T. Dunning, E. Friedman, Mahout in Action, 1st ed, Manning Publications, Greenwich, CT, USA, 2011.

[43] C.D. Manning, H. Schütze, Foundations of Natural Language Processing, 1st ed, MIT Press, Cambridge, MA, USA, 1999.

[44] A. Srivastava, M. Sahami, Text Mining, classification, clustering, and Applications, 1st ed, CRC Press, Oxford, UK, 2009.

[45] S. Biesbroek, A.D.L. Van Der, M.C.C. Brosens, et al., Identifying cardiovascular risk factor-related dietary patterns with reduced rank regression and random forest in the EPIC-NL cohort, Am. J. Clin. Nutr. 102 (2015) 146–154, doi:10.3945/ajcn.114.092288.

[46] K. Boersma, S.J. Linton, Screening to identify patients at risk: profiles of psychological risk factors for early intervention, Clin. J. Pain 21 (2005) 38–43 discussion 69-72, doi:10.1097/00002508-200501000-00005.

[47] K.F. Hulshof, M. Wedel, M.R. Löwik, et al., Clustering of dietary variables and other lifestyle factors (Dutch Nutritional Surveillance System), J. Epidemiol. Commun. Health 46 (1992) 417–424, doi:10.1136/jech.46.4.417.

[48] MacQueen JB (1967) Kmeans some methods for classification and analysis of multivariate observations. 5th Berkeley Symp Math Stat Probab 1967 1:281–297. doi:citeulike-article-id:6083430.

[49] D. Pelleg, D. Pelleg, A.W. Moore, A.W. Moore, X-means: Extending K-means with efficient estimation of the number of clusters, in: Proc Seventeenth Int Conf Mach Learn table contents, 2000, pp. 727–734, doi:10.1007/3-540-44491-2_3.

[50] P. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, doi:10.1016/0377-0427(87)90125-7.

[51] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3rd ed, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 2011.

[52] J. Pearl, Probabalistic reasoning in intelligent systems, Probabalistic Reason. Intell. Syst. (1988) 552.

[53] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106, doi:10.1023/A:1022643204877.

[54] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 1992.

[55] T. Pang-Ning, M. Steinbach, V. Kumar, Introduction to Data Mining, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2006.

[56] J.M. Tomczak, A. Gonczarek, Decision rules extraction from data stream in the presence of changing context for diabetes treatment, Knowl. Inf. Syst. 34 (2013) 521–546, doi:10.1007/s10115-012-0488-7.

[57] B. Chimieski, R. Fagundes, Association and classification data mining algorithms comparison over medical datasets, J. Heal Informatics 5 (2013) 44–51.
[58] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140, doi:10.1007/BF00058655.
[59] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, Mach. Learn. Proc. Thirteen Int. Conf. (ICML '96) 96 (1996) 148–156 doi:10.1.1.133.1040.
[60] L. Breiman, Random forest, Mach. Learn. 45 (1999) 1–35, doi:10.1023/A:1010933404324.
[61] V. Chawla N, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357, doi:10.1613/jair.953.
[62] R. Kohavi, Wrappers For Performance Enhancement and Oblivious Decision Graphs, Stanford University, 1996.
[63] Machine P, Tools L (2011) Practical machine learning tools and techniques.
[64] S.W. Yih, J. Goodman, G. Hulten, Learning at low false positive rates, in: Proceedings of the 3rd Conference on Email and Anti-Spam, 2006.
[65] S. Mani, W.R. Shankle, M.J. Pazzani, Differential diagnosis of dementia: a knowledge discovery and data mining (KDD) approach, Proc AMIA Annu Fall Symp, 1997.
[66] Maroco J, Silva D, Rodrigues A, et al (2011) Data mining methods in the prediction of Dementia: A a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res Notes 4. doi:10.1186/1756-0500-4-299.
[67] E.D. Wiener, J.O. Pedersen, A.S. Weigend, A neural network approach to topic spotting, Proc. SDAIR95 4th Annu. Symp. Doc. Anal. Inf. Retr. 332 (1995) 317–332.
[68] H.T. Ng, W.B. Goh, K.L. Low, Feature selection, perception learning, and a usability case study for text categorization, ACM SIGIR Forum 31 (1997) 67–73, doi:10.1145/278459.258537.
[69] Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, Proc. Fourteenth Int. Conf. Mach. Learn. (1997) 412–420, doi:10.1093/bioinformatics/bth267.
[70] D.H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259, doi:10.1016/S0893-6080(05)80023-1.