Multimedia Tools and Applications (2022) 81:14685–14703
https://doi.org/10.1007/s11042-021-10506-x

1198: ADVANCES IN MULTIMEDIA INTERACTION AND VISUALIZATION

Visualizing correlations among Parkinson biomedical data through information retrieval and machine learning techniques

Maria Frasca[1] · **Genoveffa Tortora**[1]

Received: 30 August 2020 / Revised: 3 December 2020 / Accepted: 5 January 2021 /
Published online: 27 February 2021

Abstract
In the last few years, the integration of researches in Computer Science and medical fields has made available to the scientific community an enormous amount of data, stored in databases. In this paper, we analyze the data available in the Parkinson's Progression Markers Initiative (PPMI), a comprehensive observational, multi-center study designed to identify progression biomarkers important for better treatments for Parkinson's disease. The data of PPMI participants are collected through a comprehensive battery of tests and assessments including Magnetic Resonance Imaging and DATscan imaging, collection of blood, cerebral spinal fluid, and urine samples, as well as cognitive and motor evaluations. To this aim, we propose a technique to identify a correlation between the biomedical data in the PPMI dataset for verifying the consistency of medical reports formulated during the visits and allow to correctly categorize the various patients. To correlate the information of each patient's medical report, Information Retrieval and Machine Learning techniques have been adopted, including the Latent Semantic Analysis, Text2Vec and Doc2Vec techniques. Then, patients are grouped and classified into affected or not by using clustering algorithms according to the similarity of medical reports. Finally, we have adopted a visualization system based on the D3 framework to visualize correlations among medical reports with an interactive chart, and to support the doctor in analyzing the chronological sequence of visits in order to diagnose Parkinson's disease early.

Keywords Biomedical data analysis · Health information visualization · Information retrieval · Machine learning

1 Introduction

In the last few years, the application of information technology to the medical field has led to the generation of large data sets that can be analyzed to identify the risk of disease and

✉ Maria Frasca
mfrasca@unisa.it

[1] Universita degli Studi di Salerno, Salerno, Italy

support the doctor in the diagnosis, therapy of the patient, and prevention of the disease itself.

The dataset Parkinson's Progression Markers Initiative (PPMI) is a comprehensive observational, multi-center study designed to identify Parkinson's Disease (PD) progression biomarkers important in the search for treatments of this type of illness [24]. PPMI is a "open source" reference study started in 2010 which aims at finding some new biomarkers, i.e., indicators of disease that represent missing links, important in the research of better treatments for PD. The PPMI dataset is the result of the cooperation of numerous researchers; it is constituted by a large quantity of data and samples collected and acquired by the participants, all volunteers, some of whom suffer from the disease. It is updated every 8 months and available online.[1] The search for these biomarkers is carried out by:

- Decide on standardized protocols for acquisition, transfer and analysis of clinical, imaging and biologic data that can be used by the PD research community.
- Develop a comprehensive dataset of clinical data and image and biological samples which is uniformly acquired. It can be adopted to estimate the mean rates of change and the variability around the mean of the collected data in early PD, patients prodromal PD subjects, and PD subjects with a mutation of the $LRKK$ 2, GBA or $SNCA$ genes.

In [26], the authors proposed a technique to identify a correlation between the biomedical data belonging in the PPMI dataset for verifying the consistency of medical reports formulated during the visits and allow to correctly categorize the various patients. To correlate the information of each patient's medical report, an Information Retrieval technique (named Latent Semantic Analysis) has been adopted. This technique constructs a concept space on selected patient information. This information is exploited to group and classify patients into affected or not by using clustering algorithms. Results revealed that the proposed technique reached 95% of effectiveness in the classification of patients.

However, the aforementioned paper had some limitations including:

- the tables selected for the analysis included only the patient's diagnostic reports, but there was no information regarding the patients' habits and lifestyle;
- only a lexical technique named Latent Semantic Analysis technique is used;
- it lacks a visualization system in order to make the doctor's understanding of the patient's status more immediate and simple and therefore facilitate the patient's diagnosis.

This paper is an extension of [26]. In particular, the following improvements have been added:

- Further improvement of the analysis using the PPMI dataset updated in February 2020, which includes additional information such as information regarding personal habits and lifestyle for each patient;
- The use and comparison with machine learning techniques such as Text2Vec and Doc2Vec;
- The adoption of a visualization system to support the doctor in diagnosing the chronological sequence of visits in order to diagnose Parkinson's disease early.

[1] https://www.ppmi-info.org

Information Retrieval (IR) and Machine Learning (ML) techniques have been adopted to correlate the information of each patient's medical report. In particular, *i)* Latent Semantic Analysis (LSA) technique is used to construct a concept space on patient information; *ii)* A technique based on the concept of Word Embedding, a methodology of natural language processing composed of a set of tools, language models, and learning techniques that allow the representation of documents through the use of vectors with real components as Tex2Vec, which through a neural network and a term-document matrix creates a numerical representation of the documents and Doc2Vec a semantic machine learning technique, can be used in different methods depending on the field of application, in particular it can be used with pre-trained models. *iii)* clustering algorithms are adopted to group patients according to the descriptions of their medical reports; *iv)* a data analysis phase is performed to classify the patient groups and finally visualize the correlations to perform the diagnosis.

The paper is structured as follows: Section 2 discusses related work, Section 3 describes the PD and the PPMI dataset, while Section 4 presents the analysis process; Section 5 reports the data analysis we performed on the considered dataset and a discussion of the results; Section 6 shows the visualization of our dataset. Finally, Section 7 concludes the paper.

2 Related work

The choice to adopt IR techniques in the medical field is increasingly common and could be a common practice in the future to support medical diagnosis.Recently, these techniques have been applied to biomedical data [9, 16, 22, 23]. In particular, Gefen and Miller [16] use the LSA on medical records related to congestive heart failure to identify patterns of associations between terms of interest. Similarly, Li and Wu [22] propose the KIP software tool for the identification of topical concepts from medical documents. An advantage of these studies is that the knowledge of the diagnosis and the treatment to be applied could be kept up to date, while one of the disadvantages could be that medical documents are analyzed without considering the real meaning of the words and this could cause errors and confusion.

Mao and Chu [23] propose a phrase-based vector space model for indexing medical documents, whilst Chou and Chang [9] have developed an IR system for doctors and patients to retrieve similar medical case records or related documents from various databases, deriving the similarity between the concepts using their relationship based on knowledge and the similarity between two sentences was measured using their root overlaps and the similarity between the concepts.

While as regards ML techniques, documents can be classified in three ways: unsupervised, supervised and semi-supervised methods, these techniques are widely used for the extraction of knowledge in biomedical data [2, 5–7, 19, 35]. In particular Beam et al. [3] presents a new reference methodology based on statistical power specifically designed to test incorporations of medical concepts, called cui2vec. This study, however, has limitations as most of the sources of health data are not easily shared, which limits the study to small local data sources. Finally, they provide a downloadable set of pre-trained embeds as well as an online tool for interactive exploration of the embeds.

Dynomant et al. [12] used the Doc2Vec algorithm to train models that allow you to vectorize documents on the PubMed database to analyze whether you can replace the statistical model PubMed Related Articles (pmra). This algorithm was able to link documents sharing MeSH labels in a similar way the pmra did.

Chen and Sokolova [8] they used Word2Vec and Doc2Vec unsupervised to analyze sentiment summary reports. They aimed to detect whether there is any latent prejudice towards or against a particular disease. They used SentiWordNet to establish a golden sentiment standard for data sets and evaluate the performance of the Word2Vec and Doc2Vec methods.

The visualization of the data is very important, especially in the medical field, it is used to synthesize all the information relating to a patient but in particular, the visualization of the data is used to support decisions and diagnosis [28, 33].

Especially Lesselroth et al. [21] underline some problems concerning the management of information at the point of care and propose strategies for a better visualization of data including multimedia displays, clinical dashboards, concept-oriented views, metaphor graphics and probability analysis.

Ropinski et al. [30] examine glyph-based visualization techniques that have been exploited when viewing spatial multivariate medical data. To classify these techniques, they derive a taxonomy of the properties of glyphs that is based on classification concepts established in the information display.

Blaas et al. [4] presented to highly interactive, coordinated view-based visualization approach that has been developed for dealing with multi-field medical data. This type of visualization is based on intuitive interaction techniques and integrates analysis techniques from pattern classification to guide the exploration process.

In our work, we use IR and ML techniques to identify eventual correlations between documents in order to recognize the different classes of patients based on the specific medical reports. Furthermore, we perform a semantic analysis, Text2Vec and Doc2vec tecniques on the medical report with the aim of highlighting the most characterizing keywords for Parkinson's disease. Finally, a visualization system was adopted to support diagnosis for doctors.

3 Dataset

PD is a neurodegenerative disease, belonging to the "*Movement Disorders*" category. It is originates from the degeneration of neurons in the brain that produce the neurotransmitter "$dopamine$". In the early stages of the disease, the most obvious symptoms are related to movement, and include tremors, stiffness, bradykinesia, postural instability, slowness in movement and difficulty walking. Afterwards, cognitive and behavioral problems may arise, such as dementia, depression, psychotic features, autonomic dysfunction, oculomotor abnormalities [17]. In particular, in PD the production of dopamine in the brain decreases consistently and the reduced levels of dopamine are due to the degeneration of neurons in an area called *substantia nigra* (cell loss is over 60% at the onset of symptoms). Moreover, from the marrow to the brain accumulations of a protein called alpha-synuclein begin to appear. This insoluble protein accumulates within neurons forming inclusions, called Lewy bodies [10]. The causes of PD are not yet known but, it seems that there are multiple elements that contribute to its development. These factors are mainly:

- Genetic mutations: among these, the mutation of the $LRRK$ 2, named also $PARK8$, is the most relevant. The heterozygous mutation, 2877510 $G \rightarrow A$, of this gene is the most commonly described, representing the majority of familial cases of cases of idiopathic PD [10].
- Toxic factors and work exposure such as some insecticides or herbicides.

The diagnosis of PD remains a clinical diagnosis because there are neither objective tests nor specific biochemical and neuroradiological markers. However, in the last decade one of the objectives of the research has been to improve the specificity of the classical diagnostic criteria.

The diagnosis of Parkinson's disease remains a clinical diagnosis since there is no objective test or specific biochemical and neuroradiological markers. In the last decade, however, one of the research objectives has been to improve the specificity of classical diagnostic criteria: in fact, the "United Kingdom Parkinson's disease Society Brain Bank" has proposed clinical criteria that are still widely used in clinical practice and research protocols. These diagnostic criteria establish that the sign necessary to diagnose Parkinson's disease is bradykinesia or akinesia, associated with at least one of the other so-called major signs mentioned above, i.e. muscle stiffness, tremor at rest and postural instability. These diagnostic criteria underline how clinical diagnosis is based on the combination of some "cardinal" motor signs and on the exclusion of symptoms considered "atypical" [17]. In conclusion, the symptoms of Parkinson's disease manifest themselves differently in different patients, who may experience some symptoms and not others, and also the rate at which the disease progresses varies from individual to individual. For this, the misdiagnosis rate can be relatively high.

The activity conducted by PPMI is an "open source" study, the data and samples collected and acquired by volunteer participants, affected and not by the disease, will allow the development of a database and a complete biorepository, which is currently available online and updated. every eight months. Being data collected from patients from various continents one of the main tasks of PPMI is to coordinate the management of the various data, defining a protocol for the collection and coding of data. The elaborated repository can be downloaded by accessing the portal of the PPMI site to allow the scientific community to conduct complete and exhaustive research.

The PPMI dataset is the result a clinical study based solely on observations aiming at fully evaluating significant cohorts of interest by using advanced imaging, biological sampling, clinical and behavioral assessments to identify biomarkers related to the progression of PD. The collected data may be helpful in the research of therapies to slow down or stop this progression.

The complete dataset consists of 145 files in CSV format, containing information about six macro-areas listed in the following:

- *Biospecimen*: collection of data related to clinical tests, such as blood collection, DNA and lobar puncture.
- *Imaging*: use of imaging techniques, such as Magnetic Resonance, Pet and DatScan through which it is possible to observe non-visible areas of the organism.
- *Medical History*: clinical history of patients from the first symptoms of the disease to the latest health conditions. The collection includes possible side effects of the medicines taken, results of neurological examinations, physical and so on.
- *Motor MDS-UPDRS*: collection of motor disturbance data through the use of the MDS-UPDRS scale to evaluate the stage of Parkinson's disease.
- *Non Motor Assessments*: collection of data related to cognitive and emotional-behavioral disorders.
- *Study Enrollment*: collection of conclusive data on particular studies conducted on patients.
- *Found*: collection on personal habits and lifestyles data.

Figure 1 shows the data model of the PPMI dataset, in which there are five entities:

- *Patient*: represents the set of patients participating in the study.
- *Event*: represents the set of tables that refer to the visits and analyzes to which patients are subjected.
- *Biospecimen Analysis Result*: represents the set of tables in which the analysis of the results for the controls to which the patients have undergone are present.
- *Family History*: a set of tables that describe the patient's family histories.
- *Medication*: a set of tables in which the medicines taken by patients are cataloged.

The entities described above are connected through relationships:

- *R*: represents the relationship that exists between the entities, $Event$, $Patient$, and $Biospecimen\ Analysis\ Result$, through the $PATH$, that is the unique attribute that identifies the patients;
- *HAS*: is the relationship that represents the connection between patients and their family history;
- *ASSUMES*: is a relationship that associates to each patient the medicines he takes.

4 The analysis process

In this section we present the process that allows us to find the correlation between the information on the visits and the patient's disease status, which may be: sick (PD and GENPD), healthy (HC, GENUN and SWEDD) and healthy with typical symptoms of the disease (PRODROMAL, i.e., subjects suffering from insomnia and have mutations of the $LRRK$ 2 gene). The version of the PPMI data used is updated in February 2020.

4.1 Dataset pre-processing

The preprocessing of a text is one of the key components for classifying the text, it is carried out by cleaning from widely used terms, conjunctions, adverbs, and in general the so-called "empty" words, but also the removal of additional spaces [32].

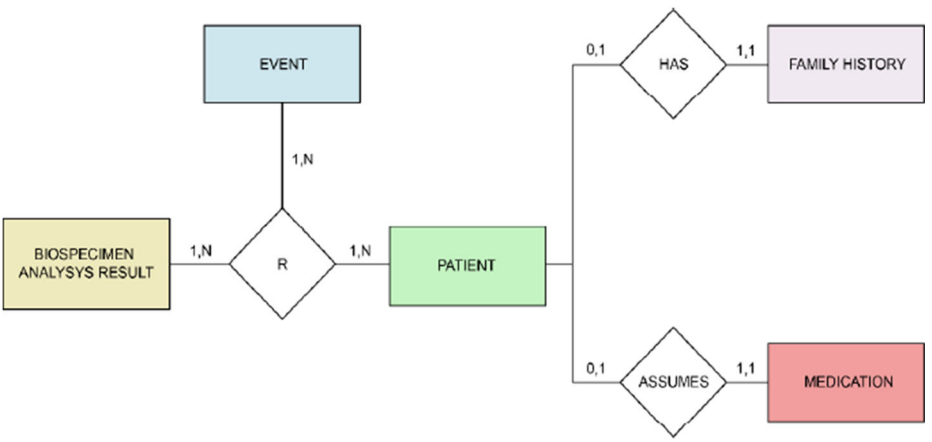

Fig. 1 The data model of the PPMI dataset

After an initial analysis of the dataset, a skimming of the tables was carried out by selecting only those of interest for the text analysis. This selection was made only on the tables in which symptoms strongly correlated with the disease appear, excluding all those containing diagnostic information. Moreover, we evaluated only the $screening\ visits$ (SC), the $basic\ line\ visits$ (BL), the $visits\ 1-15$ (V01-V015), the $symptomatic\ therapy\ visits$ (ST) and the $adverse\ events\ visits$ (LOG). The process examines them in chronological order. Following the selection and recognition phase, a modification of the tables is made, transforming in textual form only the columns related to the relevant symptoms, in which the answers of the patients to the various questionnaires or of the doctors were present in numerical form. Subsequently, the columns that did not contain any type of relevant information regarding symptoms such as numeric and text fields containing abbreviations, which may vary from table to table have been completely eliminated. Finally, a further skimming was carried out by eliminating all the tables that gave reliable information on the diagnostic status of the disease.

The process of correlating information was made sequentially as the visits progressed, adding the records of the next visit to previous ones. For each collection of documents taken into consideration, information is extracted for each individual patient. All the extracted information is kept in a new collection (Corpus). Subsequently, an initial cleaning of the text is performed. At first we made the tokenization of the text. Tokenizing a text means dividing the sequences of characters into minimal units of analysis called "tokens", after which we have reported all the text in lower case and we have removed the non-textual tokens (i.e., operators, special symbols and numbers). We have also removed white space, terms with a length of less than three characters and stopwords. In particular, we performed stopword removal by also excluding words specific of the disease, such as "parkinsonian", "parkinsonism" and "parkinson", because they could be discriminatory terms for the classification.

4.2 The correlating information process

We used a stemming algorithm to transform the words in their root form, called "theme". Then, we created the $n-by-m$ document-terms matrix A, where a generic entry $A_{i,j}$ denotes the number of times that the i_{th} term in the j_{th} document appears. For the weight associated to each pair (term, document) we used the term frequency-inverse document frequency, also known as $tf-idf$ [11]: $tf\_idf(A) = log(tf(A)) * idf(A)$. In particular, for every term t_i and document d_j in A, tf_idf is computed as follows:

$$tf\_idf(A[t_i, d_j]) = log(A[t_i, d_j] + 1) * ln\left(\frac{|d|}{\sum_{i,j} A[t_i, d_j] > 0}\right) \quad (1)$$

The tf_idf is a function used in information retrieval to measure how important a word or a document is in a corpus [27]. The tf-idf value proportionally increases the number of times a word or document appears in the corpus. In our previous work [26] we adopted LSA for correlating information. In this paper we also explore the use of two Natural languages Processing techniques: Text2Vec and Doc2Vec. The techniques used for correlating information are:

– *Tex2Vec*: Tex2Vec [31] is technique of text analysis and Natural Language Processing (NLP) by building machine learning algorithms based on text data which main goal is to provide an efficient framework with concise APIs for text analysis. It is built around

streaming APIs and iterators, which allows the construction of the corpus from iterable objects. This analysis allows us to build a matrix of document terms (DTM) and to elaborate the text by creating a map from words in a vector space. This technique is based on the concept of Word Embedding, a methodology of natural language processing to map words or phrases present in vocabulary, in a corresponding vector of real numbers, used to discover semantic correlations between them. To identify similar documents, we use cosine similarity identically. Also, in this case, a corpus of documents is built by selecting only those that coincide with the context in which the phenomena of interest reside, since even the inclusion of a large collection of high-quality documents could fail if the context of these documents does not is aligned with the phenomenon of interest. Moreover, it is necessary to have a very large corpus to create a representative sample and to increase the chances of a word appearing in it. In this case, the stemming and the SVD are not applied. To represent documents in a vector space, we need to map terms with identifiers. In such a way as to represent a set of documents as a sparse matrix, where each row corresponds to a document and each column to a term. In our case, we have created a Document Term Matrix based on vocabulary. Doing nothing but cataloging the unique terms and assigning a unique ID. After, the similarity matrix is calculated using the DTM applying the cosine similarity. The comparison between documents with the cosine similarity also takes place. These are used to find which vectors are most similar to each other and which documents have a similarity greater than a specified threshold.

– *Doc2Vec*: Doc2Vec [20] is a technique that allows you to transform textual documents into vectorial representations that protect their semantics, trying to keep all possible information expressed in the text within the vectors, for example managing to interpret the information of similarity or thematic diversity between various text blocks. In reality, the Doc2Vec is an evolution of the Word2Vec technique, which consists of a group of models, used to do word embedding, whose purpose is to translate words or sentences into vectors of real numbers, or, in a form easily computable by compilers and which it manages to represent are not the word intended as a "sequence of characters" but also the meaning it assumes, thus managing to create a coding that allows, for example, to summarize concepts of similarity or opposition about other terms. These models are nothing more than two-level neural networks trained, through an unsupervised approach, to reconstruct the linguistic contexts of words; Word2Vec takes as input a large fragment of text and builds a vector space in which each word is uniquely assigned to a corresponding vector in space. The goal of Doc2Vec is to create a numeric representation of an entire document regardless of its length.

  The purpose of Doc2Vec for its similarity to Word2Vec is to create a vectorial representation of an entire document regardless of its length, therefore, the vectors obtained will summarize the main theme or the global meaning of the entire document. It makes use of the Word2Vec model and in input, another vector is added, called DocumentID. So after training the neural network, you will have not only the word-vector (the vector representation of the words) but also a document vector (vector representation of the document). The purpose is simple, taken as input the DocumentID, the model uses the similarities between the words learned during the training (the word-vector) to build a vector that will include the words contained in it. By comparing these vectors, for example using the cosine similitude, we can then compare multiple documents with each other to verify their similarities. Doc2Vec, according to the Word2Vec approach used as

a base, is divided into two methodologies; in particular, we have the "Distributed Memory version of Paragraph Vector" (PV-DM) deriving from CBOW and the "Distributed Bag Of Words memory version of Paragraph Vector" (PV-DBOW) deriving from Skip-Gram. In particular, we based ourselves on the Word2Vec Skip-Gram, where the task of the neural network is to calculate, given an input word, the probability for each word of the vocabulary (together with all the words obtained from the training documents) to be close (juxtaposed within the text) to it. In reality, the concept of "closeness" between words is described through the definition of a measure, called windows size, which describes the number of terms to be analyzed, preceding or following the word given as input.

The network, therefore, for each input word, will have to find the probability that that specific word forms a pair with another word of the vocabulary; therefore, the net will be trained according to the number of times each pair is used. To allow training of the neural network it is necessary to provide a numerical representation of the words, as these, in the form of strings, cannot be easily used; for this reason, a vocabulary of the words obtained from the training documents is built and one-hot vectors will be used as input to the neural network; vectors of size equal to the size of the vocabulary, consisting of all negative bits except a positive one in correspondence with the reciprocal term in the vocabulary. The output of the neural network will also be a vector of the same size, as it will also use the indices to refer to the terms of the vocabulary, but, it will contain the probabilities of the various terms of being "close" to the word given in input.

4.3 The analysis of correlating information

After running the LSA, Text2Vec or Doc2Vec algorithms and obtaining the similarity matrix between the various documents, a clustering technique can be applied. In this work we use two types of techniques to compare categorization on the type of patients: the k-means [1] and the Fuzzy c-means clustering [15]. The main difference lies in the way in which the classification of the elements takes place:

- In the k-means technique the elements can belong only in mutual exclusion to a cluster and once assigned to a given cluster they can can no longer be moved. The k-means algorithm is part of the "hard clustering" techniques and we exploited the $kmeans$ function in the r package "cluster".[2] It is a partition clustering algorithm that allows to subdivide a set of objects in K groups based on their attributes, by partitioning the data set into unique homogeneous clusters whose observations are similar but different from other clusters. The k-means iteratively improves the initial centroids by minimizing the total intracluster variance, i.e., maximizing the similarity between the documents. The resulting clusters remain mutually exclusive.
- In the Fuzzy c-means technique the elements can belong simultaneously to both clusters, without any constraint. Fuzzy c-means clustering, also referred to as soft clustering or soft k-means, is used with the $fanny$ function of the r package $cluster$, each element has a set of membership coefficients corresponding to the degree of the link with a given cluster; this value can vary from 0 to 1. The Fuzzy c-means algorithm is one of the most common fuzzy clustering algorithms, the centroid of a cluster is calculated as a weighted average of all points, based on the degree of cluster membership. The

[2] https://cran.r-project.org/web/packages/cluster/cluster.pdf

clustering process is accomplished through an iterative optimization of the following function:

$$\sum_{v=1}^{nc} \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} u_{iv}^{r} u_{jv}^{r} d(e_i, e_j)}{2 \sum_{j=1}^{m} u_{jv}^{r}} \tag{2}$$

where e_i and e_j are pairs of entities selected in the set of all cluster entities. The size of this set is m, while nc is the number of clusters to identify and u_{iv} is a not negative value that specifies the membership of the entity e_i to the cluster v. The sum of all the relevances of a given entity e_i is 1, while the exponent of membership is r and can assume values between 1 and ∞. In the case r is close to 1, the behavior of the algorithm is similar to that of the k-means algorithm. The clustering process will stop when the inequality occurs [29]:

$$max_{i,v=1\ldots nc} \left| u_{iv}^{t+1} - u_{iv}^{t} \right| < \varepsilon \tag{3}$$

t indicates the maximum number of iterations, while ε represents a termination criterion. The value for ε in [0,1]. Fuzzy c-means computes a membership matrix that is used to generate clusters. We empirically set $r = 1.01$, $\varepsilon = 1e^{-20}$ and $t = 1000$.

Both algorithms needs the definition of K, the number of clusters in which the information is divided. Because we classify the patient in two groups (genetically affected, genetically not affected) we set $K = 2$. Moreover, as we have already said, in the Fuzzy c-means technique an element can belong to several clusters without restrictions and with different percentages of belonging. To overcome this problem, we have cleaned up all the spurious values, i.e., the observations that were less than $1/K$, which were therefore removed and the results obtained are shown in Fig. 2.

5 The data analysis

To perform the comparison of the processes, as we have already said previously, the techniques were performed on the dataset divided by visits. Considering that the visits follow a precise chronological order (starting from the SC-screening visit up to the LOG-diagnostic visit) an incremental subdivision of the dataset has been chosen, therefore the subset of the dataset relating to a specific visit will contain the data of it plus those of previous visits. Furthermore, we have chosen to implement a function that subdivides the dataset into a partition in which each subset corresponds to a visit and contains only the information relating to it; this is to allow future analyzes that focus solely on the data provided by a visit or by groups of visits.

The results of the clustering algorithms are analyzed, by computing Precision, Recall and F-measure for each cluster. Precision and Recall are measures used to indicate accuracy and completeness of results, respectively, while F-measure represents a trade-off between these measures. All the measures are based on a comparison between an expected result and the result obtained. In particular, Precision measures (also called positive predictive value) the ratio between the correctly obtained instances (true positives) with respect to the total number of instances returned by the processing process (true positive and false positive). The Recall measures (also known as sensitivity) the ratio of instance correctly obtained (true

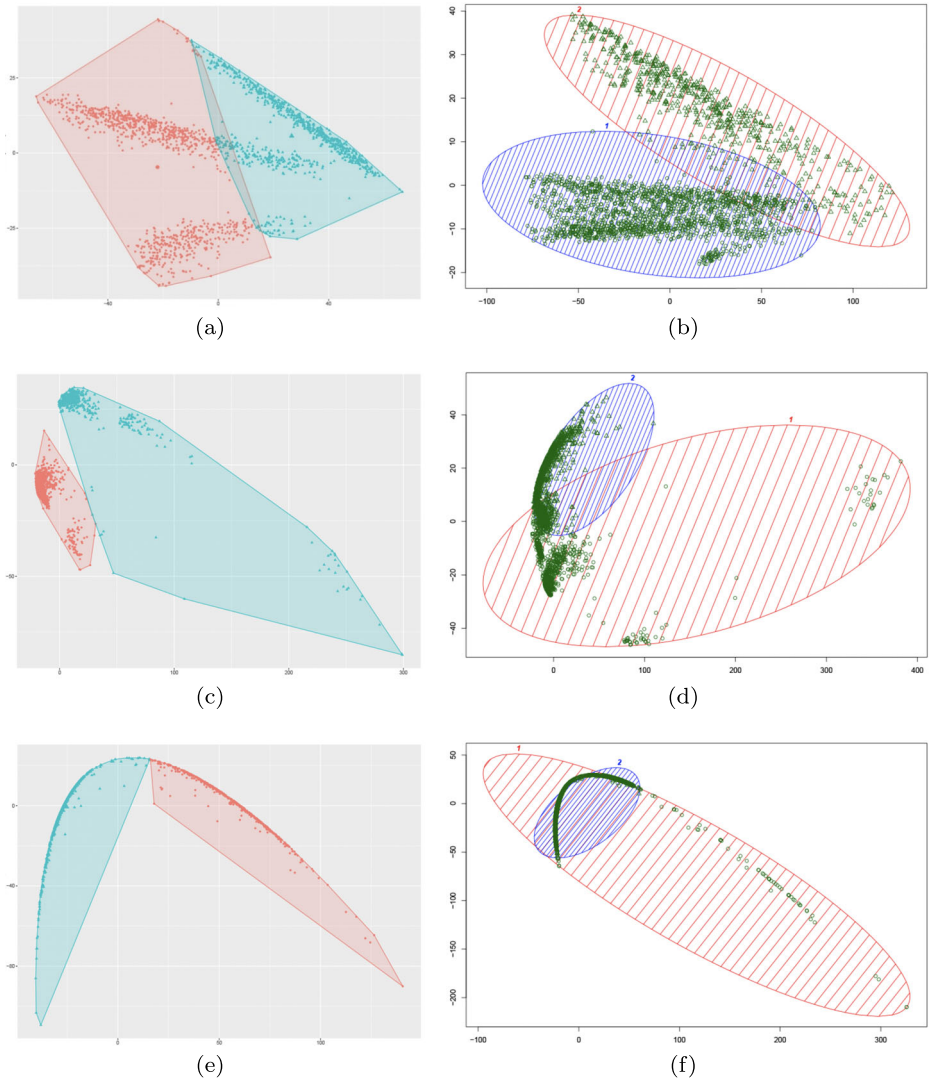

Fig. 2 Graphical representation of the clusters obtained with K-means (LSA **a**, Text2Vec **c**, Doc2Vec **e** and Fuzzy c-means (LSA **b**, Text2Vec **d**, Doc2Vec **f** algorithms for each technique

positive) with respect to the number of expected instances (true positive and true negative) [13].

$$Precision = \frac{|R \bigcap D|}{|D|}, \quad Recall = \frac{|R \bigcap D|}{|R|} \tag{4}$$

The F-measure is the harmonic mean of Precision and Recall and provides a measure of how the processing is effective.

$$F - measure = 2 * \frac{Precision * Recall}{Precsion + Recall} \tag{5}$$

where D is the list of patients returned by the classification and R is the list of patients correctly classified. The results of this analysis process are summarized in Figs. 3 and 4, where is highlighted that the best classification accuracy is reached with the fuzzy technique in each visit.

Specifically, as shown in Fig. 5, from the comparison of the results, it is shown that the techniques to which the k-means clustering is applied produce lower results than the Fuzzy clustering, except for the Doc2Vec in the SC, V02, V03 visits and the Text2Vec in the SC, BL, V01 visits where they produce better results than the other algorithms. Furthermore, despite the excellent performance of the Text2Vec k-means in the BL visit (F-Score 0,92), the technique proves to be somewhat unstable, in fact, we can see its total failure in the V04-V06 visits, where the k-means fails creating a cluster of a few dozen items. While the Text2Vec technique with Fuzzy clustering is the process that produces the best results in the series of visits from SC to V02 and from V10 to LOG. During the screening visit, both Doc2Vec (k-means) and Text2Vec (k-means and Fuzzy) produce very precise and similar classifications of patients.

Finally, we calculated the frequency of the different terms present in the documents and we have discarded the 95 quantile.

We have extracted the most relevant words from the documents related to the patient in the LOG visit, than we classified them in five categories:

- *Parental*: Patern and Sibling;
- *Symptomatic*: Pain, Sleep, Thyroid, Muscoloskeletal and Urinary;
- *Related disorders*: Hypothyroidism, Hypercholesterolemia, Diabet and Reflux;
- *Therapeutic*: Amantadin, Rytary, Arilict, Risagilin, Mirapex and Levodopa;
- *General terms*: Procedure and Full.

The words that are of most interest, come mostrato in Fig. 6 are Hypothyroidism, Hypercholesterolemia and Diabet. Regarding hypothyroidism, although no evidence of a higher frequency of hypothyroidism among patients with Parkinson's disease has been reported in the literature, there may be a concomitance between these two diseases. In fact, studies conducted on patients with PD who take levodopa / carbidopa have indicated that the reduction of TSH levels is directly related to the drug, and occurs only during the first two hours after intaking. It is not related to any significant thyroid dysfunction. This effect tends to be more discernible in males and is probably related to a primary or secondary propensity to hypothalamic levels specific for patients with PD [25]. Another study, however, has shown that the thyroid gland and its enzyme thyroperoxidase participate in the nitrosylation of serum proteins and can influence parkinsonian nitrosative stress and nitrosylation of serum alpha-synuclein, a potentially pathogenic facto [14]. Concerning hypercholesterolemia, many studies have not found a close correlation between the two pathologies, only a large prospective study [18] suggests that high total cholesterol at baseline is associated with an increased risk of Parkinson's disease. As for diabetes, a large study [34] showed that diabetes was associated with a higher future risk of PD, because the insulin receptors are expressed in the substantia nigra. The dopamine agonist bromocriptine improves glycemic control and was approved for adjunctive treatment of diabetes. Conversely, the insulin sensitizer rosiglitazone protects dopaminergic neurons in animal models of PD. It is also important to point out that both diabetes and PD are age-related chronic diseases and some pathogenic processes may underlie both conditions.

Visit	LSA K-MEANS									Patients		
	Cluster PD			Cluster GP			Results					
	PD	GP	Total	PD	GP	Total	Precision	Recall	F-measure	Total	PD	GP
SC	585	201	786	270	480	750	0,74	0,68	0,71	1536	855	681
BL	594	537	1131	463	390	853	0,53	0,56	0,54	1984	1057	927
V01	596	537	1133	461	390	851	0,53	0,56	0,54	1984	1057	927
V02	607	547	1154	450	380	830	0,53	0,57	0,55	1984	1057	927
V03	615	545	1160	442	382	824	0,53	0,58	0,55	1984	1057	927
V04	636	551	1187	421	376	797	0,54	0,60	0,57	1984	1057	927
V05	631	540	1171	426	387	813	0,54	0,60	0,57	1984	1057	927
V06	639	537	1176	418	390	808	0,54	0,60	0,57	1984	1057	927
V07	640	535	1175	417	392	809	0,54	0,61	0,57	1984	1057	927
V08	634	531	1165	423	396	819	0,54	0,60	0,57	1984	1057	927
V09	635	531	1166	422	396	818	0,54	0,60	0,57	1984	1057	927
V10	633	528	1161	424	399	823	0,55	0,60	0,57	1984	1057	927
V11	635	527	1162	422	400	822	0,55	0,60	0,57	1984	1057	927
V12	630	526	1156	427	401	828	0,54	0,60	0,57	1984	1057	927
V13	595	527	1122	462	400	862	0,53	0,56	0,55	1984	1057	927
V14	638	561	1199	419	366	785	0,53	0,60	0,57	1984	1057	927
V15	648	557	1205	409	370	779	0,54	0,61	0,57	1984	1057	927
ST	647	557	1204	410	370	780	0,54	0,61	0,57	1984	1057	927
LOG	556	457	1013	501	470	971	0,55	0,53	0,54	1984	1057	927

Visit	TEXT2VEC K-MEANS									Patients		
	Cluster PD			Cluster GP			Results					
	PD	GP	Total	PD	GP	Total	Precision	Recall	F-measure	Total	PD	GP
SC	760	205	965	96	476	572	0,79	0,89	0,83	1537	856	681
BL	960	67	1027	97	860	957	0,93	0,91	0,92	1984	1057	927
V01	749	217	966	308	710	1018	0,78	0,71	0,74	1984	1057	927
V02	757	426	1183	300	501	801	0,64	0,72	0,68	1984	1057	927
V03	765	410	1175	292	517	809	0,65	0,72	0,69	1984	1057	927
V04	1043	922	1965	14	5	19	0,53	0,99	0,69	1984	1057	927
V05	1043	922	1965	14	5	19	0,53	0,99	0,69	1984	1057	927
V06	1043	922	1965	15	5	20	0,53	0,99	0,69	1985	1058	927
V07	764	614	1378	293	313	606	0,55	0,72	0,63	1984	1057	927
V08	763	615	1378	294	312	606	0,55	0,72	0,63	1984	1057	927
V09	765	613	1378	292	314	606	0,56	0,72	0,63	1984	1057	927
V10	766	613	1379	291	314	605	0,56	0,72	0,63	1984	1057	927
V11	770	612	1382	287	315	602	0,56	0,73	0,63	1984	1057	927
V12	770	611	1381	287	316	603	0,56	0,73	0,63	1984	1057	927
V13	772	610	1382	285	317	602	0,56	0,73	0,63	1984	1057	927
V14	772	610	1382	285	317	602	0,56	0,73	0,63	1984	1057	927
V15	772	610	1382	285	317	602	0,56	0,73	0,63	1984	1057	927
ST	773	610	1383	284	317	601	0,56	0,73	0,63	1984	1057	927
LOG	766	615	1381	291	312	603	0,55	0,72	0,63	1984	1057	927

Visit	DOC2VEC K-MEANS									Patients		
	Cluster PD			Cluster GP			Results					
	PD	GP	Total	PD	GP	Total	Precision	Recall	F-measure	Total	PD	GP
SC	775	220	995	81	461	542	0,78	0,91	0,84	1537	856	681
BL	768	244	1012	289	683	972	0,76	0,73	0,74	1984	1057	927
V01	747	165	912	310	762	1072	0,82	0,71	0,76	1984	1057	927
V02	762	309	1071	295	618	913	0,71	0,72	0,72	1984	1057	927
V03	756	256	1012	301	671	972	0,75	0,72	0,73	1984	1057	927
V04	767	382	1149	290	545	835	0,67	0,73	0,70	1984	1057	927
V05	762	395	1157	295	532	827	0,66	0,72	0,69	1984	1057	927
V06	762	400	1162	295	527	822	0,66	0,72	0,69	1984	1057	927
V07	757	400	1157	300	527	827	0,65	0,72	0,68	1984	1057	927
V08	758	400	1158	299	527	826	0,65	0,72	0,68	1984	1057	927
V09	753	399	1152	304	528	832	0,65	0,71	0,68	1984	1057	927
V10	752	400	1152	305	527	832	0,65	0,71	0,68	1984	1057	927
V11	752	399	1151	305	528	833	0,65	0,71	0,68	1984	1057	927
V12	752	398	1150	305	529	834	0,65	0,71	0,68	1984	1057	927
V13	752	414	1166	305	513	818	0,64	0,71	0,68	1984	1057	927
V14	752	431	1183	305	496	801	0,64	0,71	0,67	1984	1057	927
V15	760	515	1275	297	412	709	0,60	0,72	0,65	1984	1057	927
ST	760	514	1274	297	413	710	0,60	0,72	0,65	1984	1057	927
LOG	761	513	1274	296	414	710	0,60	0,72	0,65	1984	1057	927

Fig. 3 Clustering results for the K-means algorithm and Precision, Recall and F-measure data

Visit	LSA FUZZY CLUSTERING									Patients		
	Cluster PD			Cluster GP			Results					
	PD	GP	Total	PD	GP	Total	Precision	Recall	F-measure	Total	PD	GP
SC	855	483	1338	0	198	198	0,64	1,00	0,78	1536	855	681
BL	563	239	802	494	688	1182	0,70	0,53	0,61	1984	1057	927
V01	563	241	804	494	688	1182	0,70	0,53	0,61	1986	1057	929
V02	560	239	799	497	688	1185	0,70	0,53	0,60	1984	1057	927
V03	624	454	1078	433	473	906	0,58	0,59	0,58	1984	1057	927
V04	626	448	1074	431	479	910	0,58	0,59	0,59	1984	1057	927
V05	628	443	1071	429	484	913	0,59	0,59	0,59	1984	1057	927
V06	621	432	1053	436	495	931	0,59	0,59	0,59	1984	1057	927
V07	622	430	1052	435	497	932	0,59	0,59	0,59	1984	1057	927
V08	624	433	1057	433	494	927	0,59	0,59	0,59	1984	1057	927
V09	628	436	1064	429	491	920	0,59	0,59	0,59	1984	1057	927
V10	634	447	1081	423	480	903	0,59	0,60	0,59	1984	1057	927
V11	634	447	1081	423	480	903	0,59	0,60	0,59	1984	1057	927
V12	635	450	1085	422	477	899	0,59	0,60	0,59	1984	1057	927
V13	773	640	1413	284	287	571	0,55	0,73	0,63	1984	1057	927
V14	509	315	824	548	612	1160	0,62	0,48	0,54	1984	1057	927
V15	221	93	314	836	834	1670	0,70	0,21	0,32	1984	1057	927
ST	221	93	314	836	834	1670	0,70	0,21	0,32	1984	1057	927
LOG	222	93	315	835	834	1669	0,70	0,21	0,32	1984	1057	927

Visit	TEXT2VEC FUZZY CLUSTERING									Patients		
	Cluster PD			Cluster GP			Results					
	PD	GP	Total	PD	GP	Total	Precision	Recall	F-measure	Total	PD	GP
SC	764	205	969	92	476	568	0,79	0,89	0,84	1537	856	681
BL	776	26	802	281	901	1182	0,97	0,73	0,83	1984	1057	927
V01	756	70	826	301	857	1158	0,92	0,72	0,80	1984	1057	927
V02	774	405	1179	283	522	805	0,66	0,73	0,69	1984	1057	927
V03	772	405	1177	285	522	807	0,66	0,73	0,69	1984	1057	927
V04	776	414	1190	281	513	794	0,65	0,73	0,69	1984	1057	927
V05	777	421	1198	280	506	786	0,65	0,74	0,69	1984	1057	927
V06	778	419	1197	279	508	787	0,65	0,74	0,69	1984	1057	927
V07	777	414	1191	280	513	793	0,65	0,74	0,69	1984	1057	927
V08	777	413	1190	280	514	794	0,65	0,74	0,69	1984	1057	927
V09	776	409	1185	281	518	799	0,65	0,73	0,69	1984	1057	927
V10	749	94	843	308	833	1141	0,89	0,71	0,79	1984	1057	927
V11	744	63	807	313	864	1177	0,92	0,70	0,80	1984	1057	927
V12	740	55	795	317	872	1189	0,93	0,70	0,80	1984	1057	927
V13	739	50	789	318	877	1195	0,94	0,70	0,80	1984	1057	927
V14	736	43	779	321	884	1205	0,94	0,70	0,80	1984	1057	927
V15	736	43	779	321	884	1205	0,94	0,70	0,80	1984	1057	927
ST	735	43	778	322	884	1206	0,94	0,70	0,80	1984	1057	927
LOG	735	44	779	322	883	1205	0,94	0,70	0,80	1984	1057	927

Visit	DOC2VEC FUZZY CLUSTERING									Patients		
	Cluster PD			Cluster GP			Results					
	PD	GP	Total	PD	GP	Total	Precision	Recall	F-measure	Total	PD	GP
SC	518	455	973	337	224	561	0,53	0,61	0,57	1534	855	679
BL	754	456	1210	303	471	774	0,62	0,71	0,67	1984	1057	927
V01	735	426	1161	322	501	823	0,63	0,70	0,66	1984	1057	927
V02	838	376	1214	219	551	770	0,69	0,79	0,74	1984	1057	927
V03	811	364	1175	246	563	809	0,69	0,77	0,73	1984	1057	927
V04	931	456	1387	126	471	597	0,67	0,88	0,76	1984	1057	927
V05	918	435	1353	139	492	631	0,68	0,87	0,76	1984	1057	927
V06	934	446	1380	123	481	604	0,68	0,88	0,77	1984	1057	927
V07	933	438	1371	124	489	613	0,68	0,88	0,77	1984	1057	927
V08	925	428	1353	132	499	631	0,68	0,88	0,77	1984	1057	927
V09	924	419	1343	133	508	641	0,69	0,87	0,77	1984	1057	927
V10	923	421	1344	134	506	640	0,69	0,87	0,77	1984	1057	927
V11	921	422	1343	136	505	641	0,69	0,87	0,77	1984	1057	927
V12	924	424	1348	133	504	637	0,69	0,87	0,77	1985	1057	928
V13	925	423	1348	132	504	636	0,69	0,88	0,77	1984	1057	927
V14	926	422	1348	131	505	636	0,69	0,88	0,77	1984	1057	927
V15	930	421	1351	127	506	633	0,69	0,88	0,77	1984	1057	927
ST	930	421	1351	127	506	633	0,69	0,88	0,77	1984	1057	927
LOG	934	408	1342	223	519	742	0,70	0,81	0,75	2084	1157	927

Fig. 4 Clustering results for Fuzzy c-means algorithm and Precision, Recall and F-measure data

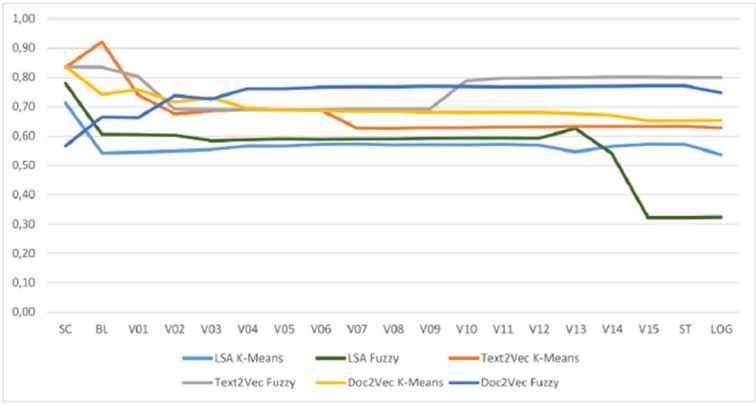

Fig. 5 Comparison among F-measures for k-means and Fuzzy c-means algorithms

5.1 Discussion

From what has been observed we can conclude that the processes with LSA produce the worst results, and generally, the results produced by the execution of the processes via K-means are lower than those with Fuzzy clustering, this is due to the intrinsic structure of the K-means algorithm that does not allow an element that can be positioned in both clusters or that it can be moved later from one cluster to another, forcing an incorrect classification of the information processed. This is because in our data there is the PRODROMAL class, which are subjects not affected by Parkinson Disease but who present symptoms characteristic of this disease, therefore a stringent clustering could lead to a higher error rate, as can be seen from the results. In conclusion, the techniques that currently produce the best results are Text2Vec-Fuzzy, in visits from SC to V02 and from V10 to LOG, and Doc2VecM-Fuzzy in visits from V03 to V09.

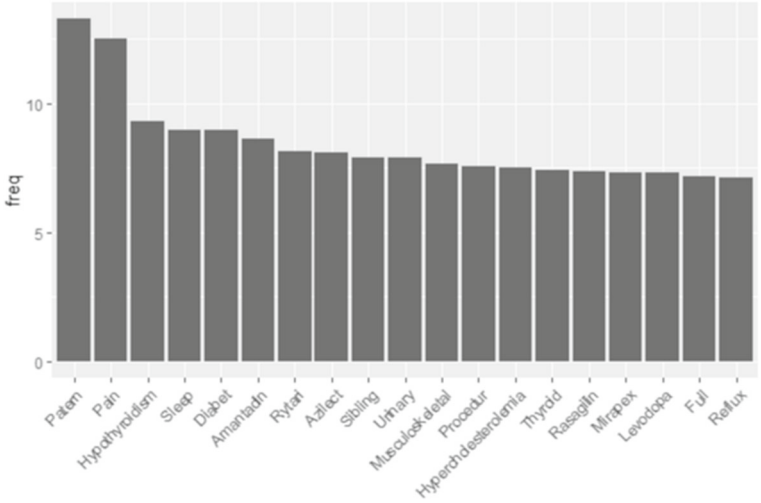

Fig. 6 Barplot representing the words with the highest frequencies in the documents related to the LOG visits

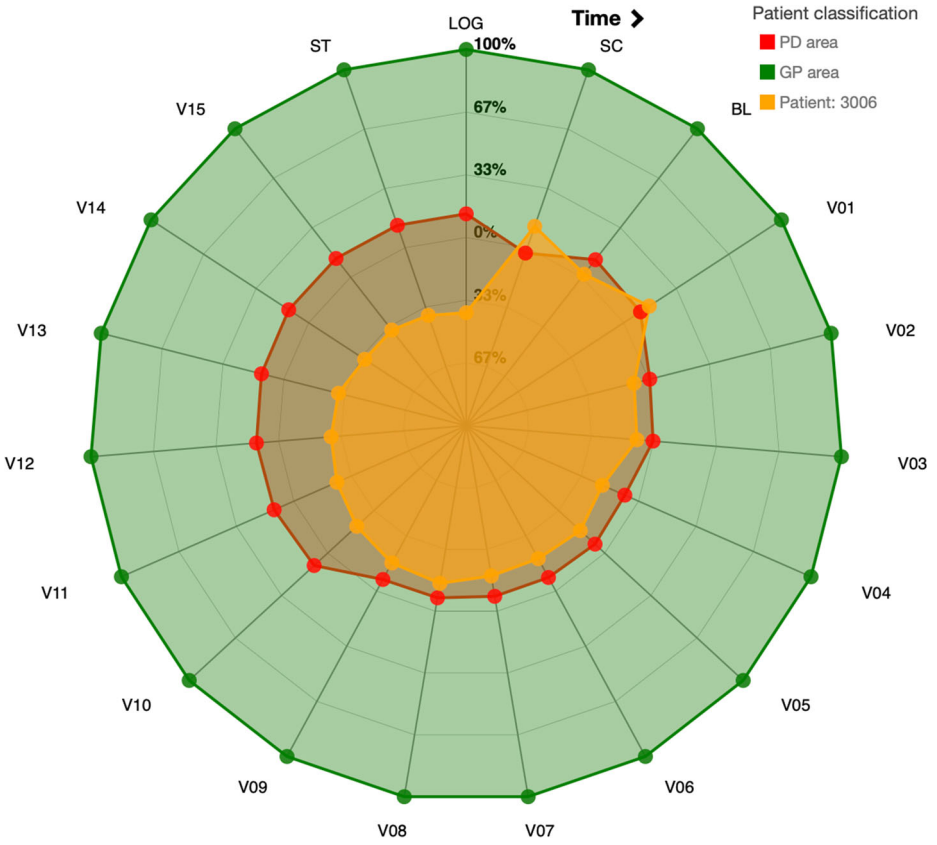

Fig. 7 A data visualization based or Radar chart for fast diagnosis

6 Data visualization

The visualization of the information is very important in each field for a personalized visualization and a better understanding of the information but especially in the medical field. For example, it can act as a decision and diagnosis support for doctors. There are many tools for viewing information and data, here we use the D3.js framework[3] that is a JavaScript library to create dynamic and interactive visualizations starting from organized data, visible through a common browser. for the visual rendering of the data, to be able to create both small tables, diagrams and statistics and complex graphic representations (including animations and other possibilities of interaction). Libraries are always linked to software that uses the functions of a programming library when a specific function of the collection is requested, which is why they only work within a program and cannot be performed independently.

In our case, we offer a visualization system, based on Radar chart, useful for the doctor to place each individual patient in one of the two clusters, in order to be able to make a quicker and faster diagnosis as the visits made in chronological order follow one another.

[3] https://d3js.org

For this visualization, we used the data obtained from both Text2Vec and Doc2Vec, based on the Fuzzy clustering efficiency obtained from the processing carried out for each visit. In fact, we used the Text2Vec for the SC, BL, V01, V02, V10-V15, ST and LOG visits and Doc2Vec for all the other visits, thus obtaining the chart shown in Fig. 7, for patient 3006. The chart is composed of three areas the green one indicates the non-sick patient, while the red one the sick patient; the orange line indicates the probability of patient 3006 to be included in the PD and GP areas. In the V02-V03 visits, the values indicate that the patient in question can be affected by Parkinson's disease. Unfortunately, this is confirmed more and more in subsequent visits.

7 Conclusion

In this paper, we proposed a techniques to identify a correlation between the biomedical data in the PPMI dataset useful to verify the consistency of medical reports formulated during the visits and, then, to correctly classify the patients into affected or not. To correlate the information of each patient medical report, Information Retrieval techniques and clustering algorithms have been adopted. Furthermore, we have created a data visualization system to support diagnosis for doctors to be able to categorize the Prodromal class of patients in particular.

In the future, the analysis of the data and the correlation between them may be further extended, in particular, it could aim to establish a precise coding standard of the dataset, so that any future developments work with the updated datasets, in which the information of the tables are described with the same dictionary. In this way, the results of the various research activities can be compared more reliably. Furthermore, in the process of textual conversion of the numerical information contained in the dataset, these should be described with medical jargon; remember that we trained a model on specific medical texts on Parkinson's, a correlation must be maintained between the two languages. Finally, we will evaluate the compression of the graphs proposed by doctors in order to validate the proposed technique.

Funding Open Access funding provided by Università degli Studi di Salerno within the CRUI-CARE Agreement.

References

1. Alsabti K, Ranka S, Singh V (2000) An efficient k-means clustering algorithm. First workshop high performance data mining
2. Anagaw A, Chang Y-L (2019) A new complement naïve bayesian approach for biomedical data classification. J Ambient Intell Human Comput 10(10):3889–3897
3. Beam AL, Kompa B, Schmaltz A, Fried I, Weber G, Palmer N, Shi X, Cai T, Kohane IS (2018) Clinical concept embeddings learned from massive sources of multimodal medical data. arXiv:1804.01486

4. Blaas J, Botha CP, Post FH (2007) Interactive visualization of multi-field medical data using linked physical and feature-space views. In: EuroVis, pp 123–130
5. Bleik S, Mishra M, Huan J, Song M (2013) Text categorization of biomedical data sets using graph kernels and a controlled vocabulary. IEEE/ACM Trans Comput Biol Bioinform 10(5):1211–1217
6. Bouadjenek MR, Verspoor K (2017) Multi-field query expansion is effective for biomedical dataset retrieval. Database 2017
7. Chen H, Fuller SS, Friedman C, Hersh W (2005) Knowledge management, data mining, and text mining in medical informatics. In: Medical informatics. Springer, New York, pp 3–33
8. Chen Q, Sokolova M (2018) Word2vec and doc2vec in unsupervised sentiment analysis of clinical discharge summaries. arXiv:1805.00352
9. Chou S, Chang W, Cheng C-Y, Jehng J-C, Chang C (2008) An information retrieval system for medical records & documents. In: 30th annual intl conf of the IEEE eng in medicine and biology sfociety. IEEE, pp 1474–1477
10. Davie CA (2008) A review of parkinson's disease. British Med Bull 86(1):109–127
11. Distante D, Risi M, Scanniello G (2010) Extending web content management systems navigation capabilities with semantic navigation maps. In: 12th IEEE Intl Symposium on Web Systems Evolution (WSE). IEEE, pp 1–5
12. Dynomant E, Darmoni SJ, Lejeune É, Kerdelhué G, Leroy J-P, Lequertier V, Canu S, Grosjean J (2019) Doc2vec on the pubmed corpus: study of a new approach to generate related articles. arXiv:1911.11698
13. Euzenat J (2007) Semantic precision and recall for ontology alignment evaluation. In: IJCAI, vol 7, pp 348–353
14. Fernández E, García-Moreno J-M, Martín de Pablos A, Chacón J (2014) May the thyroid gland and thyroperoxidase participate in nitrosylation of serum proteins and sporadic parkinson's disease?
15. Gath I, Geva AB (1989) Unsupervised optimal fuzzy clustering. IEEE Trans Pattern Anal Mach Intell 773–780
16. Gefen D, Miller J, Armstrong JK, Cornelius FH, Robertson N, Smith-McLallen A, Taylor JA (2018) Identifying patterns in medical records through latent semantic analysis. Commun ACM 61(6):72–77
17. Gelb DJ, Oliver E, Gilman S (1999) Diagnostic criteria for parkinson disease. Archiv Neurol 56(1):33–39
18. Hu G (2010) Total cholesterol and the risk of parkinson's disease: A review for some new findings. Parkinson's disease 2010
19. Khan A, Baharudin B, Lee LH, Khan K (2010) A review of machine learning algorithms for text-documents classification. J Adv Inform Technol 1(1):4–20
20. Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: International conference on machine learning, pp 1188–1196
21. Lesselroth BJ, Pieczkiewicz DS (2011) Data visualization strategies for the electronic health record. Nova Science Publishers Inc, New York
22. Li Q, Wu Y-FB (2006) Identifying important concepts from medical documents. J Biomed Inform 39(6):668–679
23. Mao W, Chu WW (2007) The phrase-based vector space model for automatic retrieval of free-text medical documents. Data Knowl Eng 61(1):76–92
24. Marek K, Jennings D, Lasch S, Siderowf A, Tanner C, Simuni T, Coffey C, Kieburtz K, Flagg E, Chowdhury S et al (2011) The parkinson progression marker initiative (PPMI). Progress Neurobiol 95(4):629–635
25. Munhoz RP, Teive HA, Troiano AR, Hauck PR, Leiva MHH, Graff H, Werneck LC (2004) Parkinson's disease and thyroid dysfunction. Parkinson Relat Disord 10(6):381–383
26. Pellecchia MT, Frasca M, Citarella AA, Risi M, Francese R, Tortora G, De Marco F (2019) Identifying correlations among biomedical data through information retrieval techniques. In: 2019 23rd international conference information visualisation (IV). IEEE, pp 269–274
27. Rajaraman A, Ullman JD (2011) Data mining. Cambridge University Press, Cambridge, pp 1–17
28. Rind A, Wang TD, Aigner W, Miksch S, Wongsuphasawat K, Plaisant C, Shneiderman B (2013) Interactive information visualization to explore and query electronic health records. Found Trends Human-Comput Interact 5(3):207–298
29. Romano S, Scanniello G, Risi M, Gravino C (2011) Clustering and lexical information support for the recovery of design pattern in source code. In: 27th IEEE Intl Conf on software maintenance (ICSM). IEEE, pp 500–503
30. Ropinski T, Oeltze S, Preim B (2011) Survey of glyph-based visualization techniques for spatial multivariate medical data. Comput Graphics 35(2):392–401
31. Selivanov D, Wang Q (2016) text2vec: Modern text mining framework for r. Computer software manual(R package version 0.4. 0). Retrieved from https://CRAN.R-project.org/package=text2vec

32. Uysal AK, Gunal S (2014) The impact of preprocessing on text classification. Inform Process Manag 50(1):104–112
33. West VL, Borland D, Hammond WE (2015) Innovative information visualization of electronic health record data: A systematic review. J Am Med Inform Assoc 22(2):330–339
34. Xu Q, Park Y, Huang X, Hollenbeck A, Blair A, Schatzkin A, Chen H (2011) Diabetes and risk of parkinson's disease. Diabetes Care 34(4):910–915
35. Zhou G, Zhang J, Su J, Shen D, Tan C (2004) Recognizing names in biomedical texts: A machine learning approach. Bioinformatics 20(7):1178–1190