Applied Soft Computing 122 (2022) 108870

Contents lists available at ScienceDirect

Applied Soft Computing

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

A fuzzy-based medical system for pattern mining in a distributed environment: Application to diagnostic and co-morbidity

Carlos Fernandez-Basso *, Karel Gutiérrez-Batista, Roberto Morcillo-Jiménez, Maria-Amparo Vila, Maria J. Martin-Bautista

Department of Computer Science and Artificial Intelligence, University of Granada, 18071, Granada, Spain

ARTICLE INFO

Article history:
Received 30 November 2021
Received in revised form 4 March 2022
Accepted 10 April 2022
Available online 26 April 2022

Keywords:
Association rules
Fuzzy logic
Data mining
Medical records

ABSTRACT

In this paper we have addressed the extraction of hidden knowledge from medical records using data mining techniques such as association rules in conjunction with fuzzy logic in a distributed environment. A significant challenge in this domain is that although there are a lot of studies devoted to analysing health data, very few focus on the understanding and interpretability of the data and the hidden patterns present within the data. A major challenge in this area is that many health data analysis studies have focussed on classification, prediction or knowledge extraction and end users find little interpretability or understanding of the results. This is due to the use of black-box algorithms or because the nature of the data is not represented correctly. This is why it is necessary to focus the analysis not only on knowledge extraction but also on the transformation and processing of the data to improve the modelling of the nature of the data. Techniques such as association rule mining and fuzzy logic help to improve the interpretability of the data and treat it with the inherent uncertainty of real-world data. To this end, we propose a system that automatically: a) pre-processes the database by transforming and adapting the data for the data mining process and enriching the data to generate more interesting patterns, b) performs the fuzzification of the medical database to represent and analyse real-world medical data with its inherent uncertainty, c) discovers interrelations and patterns amongst different features (diagnostic, hospital discharge, etc.), and d) visualizes the obtained results efficiently to facilitate the analysis and improve the interpretability of the information extracted. Our proposed system yields a significant increase in the compression and interpretability of medical data for end-users, allowing them to analyse the data correctly and make the right decisions. We present one practical case using two health-related datasets to demonstrate the feasibility of our proposal for real data.

1. Introduction

Due to the growing necessity for analysing health data properly, the development of robust systems that enable knowledge discovery from this sort of data has become a subject of great interest for companies and researchers. The extraction of hidden knowledge from health-related data, diagnostic and co-morbidity [1] analysis to create valuable medical information and improve healthcare services have all become vital challenges for health management and medical decision support [2,3].

Another significant challenge in this domain is the interpretability of the data mining systems. Interpretability can be defined as the degree to which a human can understand the cause of a decision [4] or the degree to which a human can consistently predict the results of the model [5]. The greater the interpretability of a system is, the easier it is for end-users to understand why certain decisions or predictions have been made.

In order to improve the interpretability and diagnostic and co-morbidity analysis of the medical data, we can use valuable techniques such as association rules extraction and fuzzy logic. The first is a popular unsupervised approach used to explore and interpret large transactional datasets to identify unique patterns and rules [6]. The latter was introduced in [7] and allows real-world data to be represented and analysed in a better way. In other words, the fuzzy set theory provides efficient mechanisms to manage incomplete or imprecise information.

Association rules have been widely used for pattern discovery in different domains [8–11]. Specifically, in medicine, we can find several studies that use association rules to extract hidden patterns which are useful for the diagnosis and subsequent possible treatment of a patient [12–14]. Taking into account the above and the fact that in medical databases there are incomplete and

* Corresponding author.
E-mail address: cjferba@decsai.ugr.es (C. Fernandez-Basso).

https://doi.org/10.1016/j.asoc.2022.108870

imprecise data (mostly continuous features), approaches such as fuzzy association rules would seem to be the most suitable for being an excellent solution to discretize numerical values softly and uncover and represent hidden relationships in an understandable way for end-users [15].

Another important aspect is that standard association rule mining (ARM) algorithms do not solve multitasking rule problems, because they ignore the correlation between tasks. There are algorithms such as multi-task association rule miner (MTARM) that works in a way that joins several rules to deal with this type of problem. It divides the rules into individual tasks and through a voting system generates new rules that produce global results called multitask rules [9]. In our study we approach the problem in a distributed way by dividing the problem, finally unifying the different solutions.

Let us consider the following example: we have a medical-related dataset consisting of the attributes *diagnosis*, *occupation*, and *age* (where each instance in the dataset corresponds to a patient's surveillance), and we want to know the number of patients with a specific diagnosis in a given range of ages.

We can solve the problem by means of a simple query. This approach is helpful if it is intended for performing a shallow analysis of the data. However, it does not allow a more detailed analysis, such as determining the number of *young* patients with a *complex* diagnosis. In this situation, technologies such as fuzzy association rules are revealed as the most appropriate technologies to help with these problems.

One of the main goals of medical databases is to store historical data about the patients. That is why when using this sort of database, we should keep in mind that the use of data mining methods runs into problems when they are used to analyse vast amounts of data and become less efficient at processing and analysis. To tackle this problem, we must implement the algorithm in a distributed environment.

In this study, we propose creating a fuzzy-based medical system for pattern mining in a distributed environment. The system allows end-users to discover and interpret hidden patterns from health-related data, thus facilitating diagnostic and co-morbidity analysis, subsequent treatment, and also the early prevention of potential diseases. Our proposal is based on tools such as association rules, fuzzy logic, and Big Data. The proposal is entirely automatic and unsupervised, allowing us to experiment with both labelled and unlabelled data. We now summarize the main contributions of this paper:

- **Data enrichment -** During this process, the original medical database is transformed and adapted (features engineering) to prepare the data for the data mining process to generate more interesting patterns.
- **Feature fuzzification -** Through this process, we perform the fuzzification of the features enriched in the above step, enabling us to treat imprecise data and discover and analyse relevant patterns and relationships.
- **Visualization -** Finally, we visualize the obtained results efficiently and in a user-friendly way to facilitate the analysis and improve the interpretability of the information extracted.

The rest of the paper is organized as follows: Section 2 summarizes the main work developed in this research area. Section 3 presents the proposed fuzzy-based system for pattern mining in Big Data. Section 4 presents one case from real-world medical data and we discuss the obtained results. Finally, in Section 5, the conclusions derived from the analysis are presented.

2. Related studies

As mentioned above, in this paper, we propose a fuzzy information-based system for discovering hidden patterns and relationships to address the problem of interpretability and diagnostic and co-morbidity analysis of medical-related data. One of the main problems of such studies is how the results are presented to end-users. Most studies show results without focusing on end-user understanding. Therefore, the main objective of this research is to improve the interpretability of obtained results in medical data centres using data mining techniques such as association rule discovery and fuzzy logic. This section presents research on data mining and fuzzy-based systems, mainly focused on the medical domain.

2.1. Data mining system

Data mining techniques have been widely applied in numerous fields of science. An instance where we can observe this is in energy [16–19]. In [16,17] the authors review how some traditional data mining techniques have been used to obtain construction-related information. In social science, we can observe numerous studies about data mining [20–22] where text pre-processing is applied. Other fields such as Physics and Astronomy apply pre-processing techniques to images [23–26]. As in our case, data mining techniques are also widely used in the medical field, as we can observe in [27–29]. In this field, we can classify the studies into two groups, those dealing with imaging data [30–33] and those dealing with medical records [34–37].

In a more specific health study [28], such as the analysis of brain signals, problems such as sample size and signals considered as noise are encountered. In this study, the authors present a plausible method to detect and distinguish directions from Electroencephalography (EEG) signals by using feature extraction techniques to perform brain signal processing.

In this other study [30], an automated system for the extraction and classification of tumours from magnetic resonance images has been developed. The proposed system consists of five main steps: tumour contrast, tumour extraction, multi-model feature extraction, feature extraction and classification. Other techniques used in some studies, such as [38–41], focus on unsupervised algorithms.

We can observe studies that pre-process the source data to improve data quality and improve results. Different types of data require different processing technologies. Most structured data usually require classical pre-processing technologies, such as data cleaning, data integration, data transformation and data reduction [34,42,43].

Big data technology has many areas of application in the healthcare sector [27,44–46], such as predictive modelling and clinical decision support, disease surveillance and research. Big data analytic often leverages analytical methods developed in data mining, such as classification, clustering and regression.

In our study we have proposed a series of techniques that differ from other studies in the way we apply the techniques and unify them so that at the end of the procedure we have a set of data ready for knowledge extraction and interpretation.

We have enriched our data by adding the different diagnoses that occur during the duration of the patient's stay in the medical action protocol to resolve the proposed diagnosis using external sources. We have used basic data pre-processing techniques, such as the detection and elimination of missing values within the medical data centre, as well as the elimination of outliers by selecting fields not relevant to our study.

The next step in our procedure is the transformation of our data by applying fuzzy techniques to finally create association rules with the different data in order to detect the existing relationships between the different fields that we are going to analyse in our study.

2.2. Fuzzy-based system

Most of the problems presented by the medical data centres are related to the way to structure the information within the database, as well as the way to represent it, so that it can be correctly interpreted by the end user. Therefore, it is necessary to use fuzzy techniques to transform imprecise data into accurate data capable of being interpreted by the end-user. In [47] the authors propose new measures of accuracy and usefulness for fuzzy association rules extraction from medical relational databases. The approach presented by the authors allows the significant reduction of the number of rules without information being lost.

Another interesting example found in the state of the art is the following study [48] aimed at the application of fuzzy techniques in the healthcare sector. It shows how wearable sensors can be used to create a system for recommending specific prescriptions for patients with diabetes.

There is also a theoretical study [49,50] on solving linearly posed problems and applying a fuzzy programming algorithm, where the results can be obtained in a way that eliminates uncertainty.

Such stacks can be made fully scalable by applying the algorithms in a distributive manner [51]. This study presents biomedical data stored in the cloud and demonstrates how such algorithms are ideal for solving large-scale problems.

In our study, we look at the fuzzification of diagnoses and how through the co-morbidity of some diagnoses we can give results on patients so that they are displayed through our application as simple, medium or complex, depending on the number of co-morbidity diagnoses they present, thus allowing us to represent a traceability of the patient's history.

2.3. Co-morbidity analysis

Another aspect that we address in the scope of this study is to take into account the co-morbidity of the different diagnoses. Co-morbidity is the occurrence of a disease as a function of having a previous disease in the same person [52]. At the same time these co-morbidity data are often reported statistically, mainly in the context of academic research to inform the health system and public health agencies in their decision-making.

There are very few studies that focus on the analysis of data directed to the initial diagnosis as well as, the diagnosis of some co-morbidity diseases, generated by the main diagnosis. This study [35] focuses on autism in children and on the different degrees according to the different co-morbidities they suffer from. In another study [53] we found in the state of the art is aimed at detecting unnecessary blood tests, giving as an example patients with upper gastrointestinal bleeding and patients with unspecified bleeding in the gastrointestinal tract in order to analyse the amount of calcium and haemocytes in the blood. Two experiments are performed, firstly, labelling the different tests that are promising and secondly grouping patients with co-morbidity.

Finally, a case study is shown on the disease called diabetic retinopathy, which is a co-morbidity generated by diabetes [54], and how through the application of classification techniques together with fuzzy techniques and balancers we can determine which patients are most at risk of suffering from this type of co-morbidity.

Part of our study has focused on the analysis of the co-morbidity of our records. Based on this analysis, we will be able to interpret patients as simple, standard and complex, by applying fuzzy logic.

3. A fuzzy-based system for pattern mining in Big Data

In this paper, we propose a complete system for managing and extracting information in health systems using big data architecture. It presents a process of integration, data processing with two novel phases of enrichment and treatment of uncertainty in this type of data using fuzzy techniques. The system also integrates an algorithm for extracting fuzzy association rules in Spark and tools for visualizing and interpreting the results by end-users.

3.1. System architecture and workflow

Fig. 1 depicts the complete system that follows our proposal. It can be divided into three big blocks. In the first, the data collection is carried out. Hospital data in Spain follow a standard format concerning basic patient data for each visit to a hospital, a diagnostic testing centre, surgical interventions or specialist consultations. This data can then be aggregated from external sources from other areas or services of the hospital depending on the hospital and its data management system. Therefore, depending on the type of source, be it a hospital, medical tests from one of the areas of the hospital or databases from each of the medical departments of the hospital, our system collects the data and merges it with the patient's historical data in order to have them all modelled according to the patient. In this way, for example, in the case of patients with recurrent visits to different services and departments of the hospital, we can see their traceability and diagnosis in each of these visits. In addition, this innovative system can merge hospital data into a single database with all the validations described above in this first phase of the data collection process.

The data is then stored in Big Data architecture that allows large data sets from heterogeneous sources to be used. For this purpose, NoSQL databases [55,56] have been used because they provide great flexibility for storing data from heterogeneous sources and large volumes. On the other hand, distributed processing tools such as Apache Spark [57] have been used to manage, process and analyse this massive data efficiently. Using this type of technology, the system processes this data through three modules: preprocessing, enrichment and fuzzification, which will be explained in Section 3.3.

In the last block, knowledge extraction from the processed database will be carried out. This block uses an algorithm of fuzzy association rules in Big Data [58] which will allow us to extract association rules using Spark. In addition, our user interface will implement some of the most widely used tools for visualizing association rules to facilitate the understanding and interpretability of the results to the end-users.

This whole process has been applied in the scope of the BigDataMED project to different hospitals in Spain. Here we will present the results obtained from 2016 to 2020 in two hospitals, one in Marbella and the other in Granada. However, the proposed system is general and can be applied to other hospitals and medical centres.

3.2. Data collection

The developed workflow has been used in two Spanish hospitals, the clinical hospital of Granada and the hospital of Marbella. The figure below depicts the different data sources collected and added to the database (see Fig. 2).

Every piece of data has been collected through different procedures. On the one hand, we have considered the data provided by the medical management systems, which, by means of user applications, collect the data from the different services such as consultations, emergencies, etc. On the other hand, we have

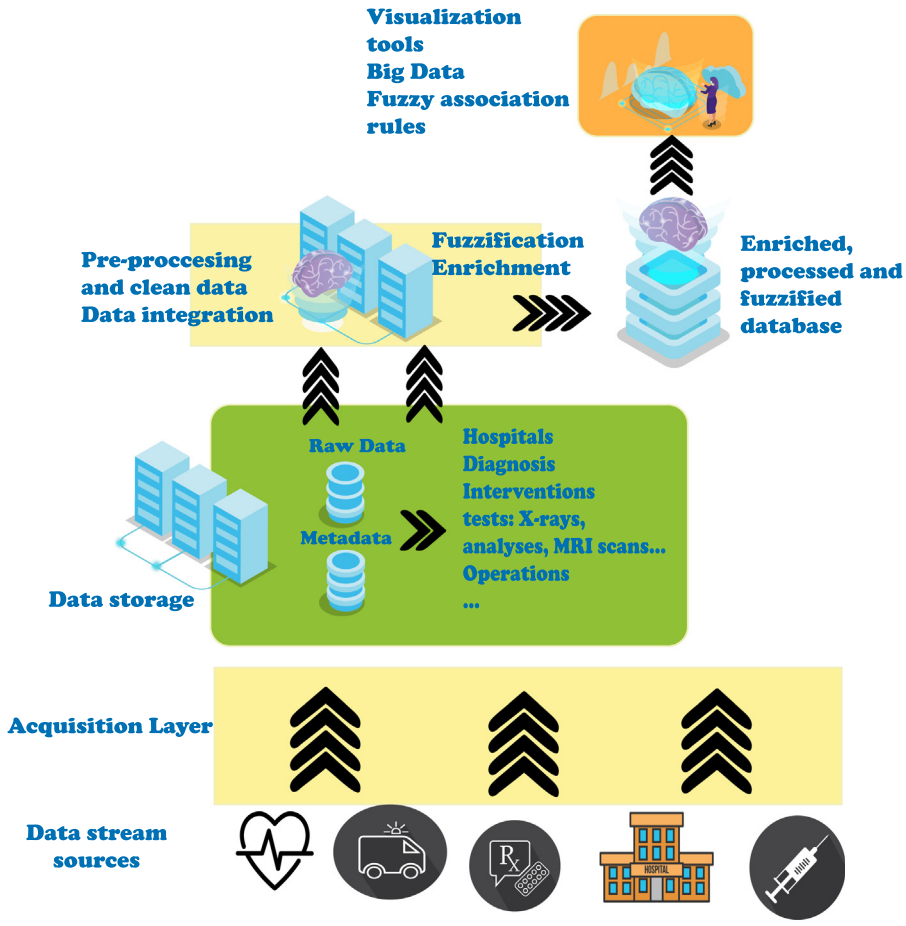

Fig. 1. General process of our proposal.

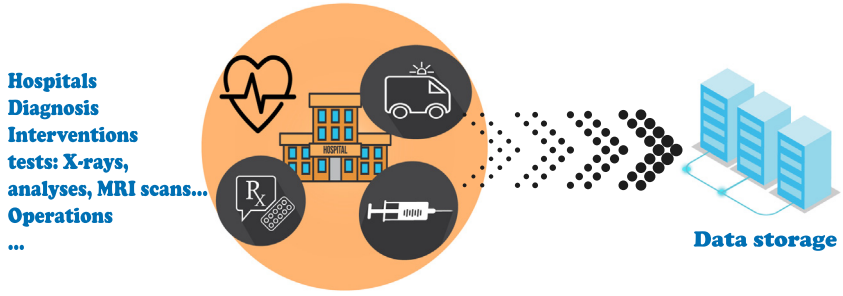

Fig. 2. Diagram of the type of data collected for the hospital system.

the data sources of the surgical interventions, diagnostic tests, etc. In particular, all these data have to be merged using the patient's history and adapting the fields so that these data can be merged by modelling their relationships. For example, the patient has to be related to the diagnostic tests performed, the surgical interventions and other sources in the system.

3.3. Pre-processing, enrichment and fuzzification

One of the essential phases in the extraction of interesting knowledge from large datasets is the preparation of these datasets for the application of data mining techniques. This phase is called data pre-processing and encompasses different groups of data processing techniques that enable the creation of a dataset with better conditions for knowledge extraction.

In our proposal, several pre-processing modules are presented. In them, some variables have been transformed to better model the information they contain. Moreover, in some cases, external data have been used to enrich the data contained in the dataset variables; finally, an uncertainty treatment process has been carried out using fuzzy logic. In this process, a novel fuzzification process has been used that either automatically or guided by expert knowledge can use fuzzy labels to better model the information and allow more interpretative results for the end-users.

3.3.1. Pre-processing data

In the first pre-processing module a classical preprocessing is carried out, eliminating codes from the database that do not contain useful information, normalizing some values and transforming some factor variables. Subsequently, processing which is more oriented to medical data has been carried out. Variables such as history codes, postcodes or dates have been processed into data that better represent the information. For example, postcodes have been transformed into data of municipalities,

Table 1
Example of the processing of temporary data from the hospital database.

Date admission	Date discharge	Birthday
13/5/2016 15:14:30	20/5/2016 15:14:30	11/2/1957
10/10/2019 15:16:45	12/10/2019 15:16:45	11/2/2016
↓		

Admission time	Season	Age	Patient type
7	Spring	64	Adult
2	Autumn	5	Child

Table 2
Element description.

Features	Codification	Method for enrichment
Diagnoses	International Classification of Diseases (ICD)	External API
Origin of patients	Spanish health system	Database
Reason for discharge	International Classification of Diseases (ICD)	External API
Reason for admission	International Classification of Diseases (ICD)	External API
Surgical procedures	International Classification of Procedures	External API
Diagnostic tests	International Classification of Test	External API
Services/ departments	Spanish health system (SNS)	Database
Other data[a]	Andalusian health system (SAS)	Database

[a]Rest of the variables related to the management and information of the Spanish hospital system.

cities and countries and dates by their month, year and day of the week, and whether the day is a holiday. In addition, the dates have been processed with the patient's age and data related to the patient's stay in hospital, such as the time of admission, stay in the Intensive Care Unit (ICU), etc. This process can be seen in the example in Table 1.

3.3.2. Enrichment

Some of the data collected within the database come from taxonomies, external coding dictionaries. This is the case for diagnoses, the origin of the patients, reasons for discharge or admission, surgical procedures and 22 other variables. It is for these types of variables that within our processing workflow, we have created a data enrichment module in which we have added information to the codes associated with each of these fields, or we have modelled their structure in order to add the information found in the external taxonomies and databases where the codes are decoded. Table 2 shows some of the characteristics that required the extraction of external information to improve their interpretability and meaning.

For this enrichment, a series of processes have been created to extract the information contained in the coding of the variable and this information has been modelled to add to the database. This occurs because the coded variables such as diagnoses have a tree-like coding that includes other levels. In addition to this, at the lower level, we find interesting information such as synonyms, alterations or problems related to the disease. Therefore our pre-processing for each diagnosis will generate up to 5 different levels of the disease. It will also generate information such as:

- Applicable to these diseases (list of diseases).
- Approximate synonyms.
- The disease is grouped within the Diagnostic Related Group (list of diseases).

An example of this process can be seen in Table 3, in which we can see how we process the diagnosis in question, as well as how the different new elements providing more information about it are grouped.

This process has been created for the different variables in the table. For each, the type of data and the type of information of interest were taken into account for their extraction.

3.3.3. Fuzzification

In the final part of the processing, we have developed a novel feature fuzzification methodology to improve data interpretability. This last phase is crucial as much of the data coming from the system is challenging to represent and interpret by end-users, as it is a continuous value with measurements that are often complex to interpret and understand. Fuzzification of these data can improve the results found by the mining algorithms, and at the same time, increase the interpretability of the obtained results.

We propose a fuzzification algorithm that allows an automatic treatment of data values according to their distribution, depending on the variables in the dataset or information provided by an expert (see the types of input provided to the algorithm). For this purpose, we have developed a distributed algorithm in Spark following the MapReduce philosophy. This allows us to process large amounts of data, such as in the case of the data stored by different hospitals in the Spanish health system.

The general process is described in Algorithm 1. For this, we use Spark for the distribution of data along the cluster. The algorithm has input a dataset, a python dictionary (hash) and an integer. The dictionary is used to store the ranges and labels of the variables that the experts have defined. On the other hand, the default number of labels is used for variables that the users have not defined and will be created automatically based on their distribution. If the values depend on dataset variables, we will pass a dictionary containing the labels and the validation function applied to the dataset variable to the algorithm. For example, if we depend on the number of diagnoses, we will pass the variable number of diagnoses and the function that returns a label with its membership value.

The whole process will be processed in a distributed way. It should be noted that Spark automatically divides the data into chunks for the distributed computation. We have specified this with the acronym DCS (distribute computing using Spark) and representing each chunk of data by S_i. In line 6, we can see that a global variable is used throughout the cluster, which is then used by the function that distributes the computation through MapReduce (line 8 of Algorithm 1).

Additionally, in line 8, the procedure calls to the *fuzzification* function described in Algorithm 2. This function is divided into different parts. Firstly, it checks if the name of the variable is found in the *Intervals* hash-list, if it is found in the python dictionary, the new fuzzified variables are created using the names of the labels specified by *Intervals* and its configuration (i.e. computation of membership degrees) attending to the specified interval (see lines 10–16 of Algorithm 2). If the variable is not found in the dictionary, an automatic procedure is used that divides the values of the variable into several intervals defined in *DefaultIntervals* according to the percentiles of the variable.

Fig. 3 shows an example with the value of $DefaultIntervals = 3$ where the y-axis represents the degree of membership and

Table 3
Example of the processing of a diagnosis C00.4.

Diagnosis					
C00.4					
			↓		
Dig level1	Dig level2	Dig level3	Application to	Approximate synonyms	Group diagnosis
Neoplasms	Malignant neoplasms of lip, oral cavity and pharynx	Malignant neoplasm of lip	Cancer, lower lip, inner aspect	Malignant neoplasm of frenulum of lower lip ...	Tracheostomy for face, mouth and neck diagnoses or laryngectomy with mcc

Algorithm 1 Main Spark procedure for fuzzification preprocessing algorithm.

Input: *Data:* RDD transactions: $\{t_1, \ldots, t_n\}$
Input: *DefaultIntervals:* number of intervals automatically generated by the algorithm
Input: *Intervals:* Hash-list of intervals for each variable: $\{Variable_i : [\{Intervals\}, \{Labels\}], \ldots, Variable_p : [\{Intervals\}, \{Labels\}]\}$
Output: Fuzzy transactions containing fuzzified values

Start Algorithm

Features = Dataset.NameFeatures()
broadcast(*Global_Features*) #Create a broadcast variable for its use across the cluster
DCS in q **chunks of Data:** $\{S_1, \ldots, S_q\}$
    $FuzzyData_{S_i} \leftarrow S_i$.**Map** ($Fuzzification(t_k \in S_i)$)
    # Map function computes independently each transaction in S_i
    $FuzzyDatabase =$
    = **ReduceByKey**($Aggregation(FuzzyData_{S_1}, \ldots, FuzzyData_{S_q})$)
return *FuzzyDatabase*

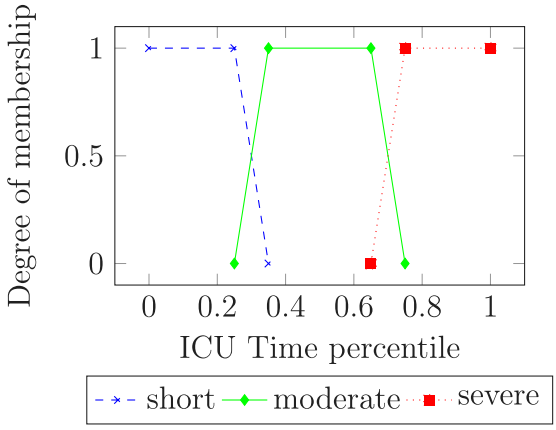

Fig. 3. Example of automatic execution with 3 default intervals.

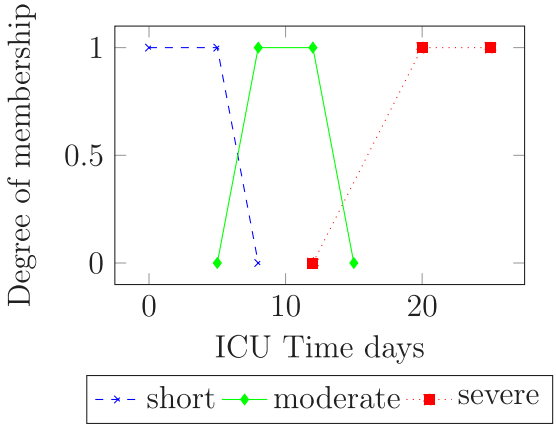

Fig. 4. Example of expert definition execution with 3 intervals.

x-axis the percentile of the variable. In this example, the percentiles used have been 25 and 37,5 for defining the trapezoidal form of the first label and left part of the second label, and 62,5 and 75 for defining the right part of second label and the third label. So, the *GenerateIntervals* function divides the set into k equidistributed fuzzy sets using the corresponding percentiles. For instance, for $k = 4$, the considered percentiles are computed as follows: $\{\frac{100}{k+1}, \frac{100}{k+1} + \frac{100}{(k+1)(k-1)}, \frac{2 \cdot 100}{k+1} + \frac{100}{(k+1)(k-1)}, \frac{2 \cdot 100}{k+1} + \frac{2 \cdot 100}{(k+1)(k-1)}, \frac{3 \cdot 100}{k+1} + \frac{2 \cdot 100}{(k+1)(k-1)}, \frac{3 \cdot 100}{k+1} + \frac{3 \cdot 100}{(k+1)(k-1)}\}$ which results in $\{p_{20}, p_{26.6}, p_{46.6}, p_{53.3}, p_{73.3}, p_{80}\}$. On the contrary, the *FuzzyDivision* function uses the defined intervals of the global variable *Intervals*.

This processing method (Algorithms 1 and 2) has a complexity $O(n/c)$ where n is the number of transactions and c is the number of computation units used in a distributed way. This complexity is due to the fact that the algorithm must go through the data set elements transforming them into the different fuzzy labels.

algorithm 2 Fuzzification function.

Input: *Data:* A transaction: $t_k = \{item_1, \ldots, item_m\}$
Global distributed variable: *Intervals:* Hash-list of intervals for each variable : $\{Variable_1 : [\{Intervals\}, \{Labels\}], \ldots, Variable_p : [\{Intervals\}, \{Labels\}]\}$
Input: *DefaultIntervals:* number of intervals automatically generated by the algorithm
Output: Fuzzy transaction

Start Algorithm

Features = Dataset.NameFeatures()
DistributeVariable(Features)
DCS in q **chunks of Data:** $\{S_1, \ldots, S_q\}$
i=0
do
    # Check if the variable exists in the hash list
    if Feature[i] ∈ Intervals **then**
        Interval=Intervals[Feature[i][0]]
        Labels=Intervals[Feature[i][1]]
    else
        Interval = GenerateIntervals(DefaultIntervals,Data[Feature[i]])
        Labels = GenerateLabels(DefaultIntervals)
    end if
    for $j = 0; j < |Labels|; j$++ **do**
        FuzzyData[Label]=FuzzyDivision(Interval[j], Interval[j+1], type)
        # type ="linear", "exponential", "logarithmic"...
        i++
    end for
while $|Feature| > i$
return *FuzzyData*

For the use case under study, the experts determined different intervals for generating the fuzzy labels. These depend on the nature of the variable, e.g. the ICU stay could be defined automatically as we have seen in Fig. 3 or defined by an expert as in Fig. 4.

3.4. Data mining: Fuzzy association rules in Big Data

After data pre-processing and fuzzification, data mining techniques were applied to the processed data. In particular, an algorithm for association rule mining was applied in Big Data

Table 4
Different data sources.

Dataset	Documents	Features
Marbella hospital	750 000	273
Granada hospital	220 000	273

(BDFARE Apriori-TID Big Data Fuzzy Association Rules Extraction [58,59]). This algorithm has also been implemented following the MapReduce paradigm under the Spark Framework. It enables the processing of huge sets of fuzzy transactions, finding frequent itemsets and fuzzy association rules exceeding the imposed thresholds for support and confidence, given a set of α-cuts.

4. Use case: medical records

The results obtained by applying our proposal should be analysed from two points of view. On the one hand, the efficiency and capacity of our distributed processing and management system that allows the processing of large datasets from different sources and with heterogeneous and massive data in a more efficient way. On the other hand, the fuzzification of the variables and their application to discover fuzzy association rules through this processing.

The aim is to extract patterns between diagnoses and patient characteristics to study co-morbidity and improve our knowledge about the information in our system. This can improve tasks such as obtaining a diagnosis or preventing a possible diagnosis related to patient factors such as obesity, other diagnoses, smoking, etc.

4.1. Data sources

In order to validate the whole process and the architecture presented above, a use case will be carried out by extracting fuzzy association rules from two hospitals in the south of Spain. For this purpose, the data used have been collected from the different systems of each of the hospitals (emergency, hospitalization, consultations, etc.). We can see the characteristics of each of them in Table 4. The table already shows the values of the records and the characteristics of the database before preprocessing.

All these data have been collected automatically from the different hospitals. They have also been stored in our system explained in Section 3.1. By having these data in our tool, we have been able to carry out the processing and knowledge extraction processes.

4.2. Data transformation and enrichment

Having all the raw data in our architecture, we will explain how the knowledge extraction process would be. This will be done using the data explained above and analysing the relationships that can be obtained from the dataset after pre-processing, cleaning and enrichment of the data. We have aimed to study the co-morbidity within the different hospital data and the complete dataset. For this purpose, it has been necessary to carry out the steps described in Section 3.3.

In the first step, we have obtained the data from the hospitals and transformed all time-related variables as explained in Section 3.3.1 and the example in Table 1. In addition, some other variables have been processed by normalizing or changing their formats to improve the performance of the algorithms.

In the next step, the enrichment phase has been applied, obtaining information from external sources such as those discussed in Table 2 and an example can be seen in Table 3. This is necessary because all databases come coded with codes from different external sources. Because of this, the characteristics of the datasets have to be extracted through external APIs, which has allowed us to add information related to diagnoses, interventions, elements such as hospitals, nationalities etc., thus improving the information contained in the system. An example can be seen with the diagnostic variable of the execution of this type of process in Table 3.

Table 5
Features after processing phase.

Feature	Group	Origin
Age	User data	Pre-prossessing phase
Origin	User data	Enrichment phase
Health (public or private)	User data	Raw data
...	User data	17 features more
Name of hospital	Hospital data	Raw data
Department	Hospital data	Enrichment phase
Type of service	Hospital data	Enrichment phase
Diagnosis (main and 19 supplementary)	Hospital admission	Enrichment phase
Admission service	Hospital admission	Enrichment phase
Reason for discharge	Hospital admission	Enrichment phase
Type of diagnosis	Hospital admission	Enrichment phase
Type of diagnosis(1 per diagnosis)	Hospital admission	Enrichment phase
Disgnostic factors(3 per diagnosis)	Hospital admission	Enrichment phase
...	Hospital admission	56 features more

The dataset obtained from the set has various variables that can be grouped into three main groups. These would be the user's characteristics such as age, origin, history, etc. Others would be the information corresponding to the hospital and its services such as name, location, service, department, etc. Finally, we would have the information related to the entry of the patient in the system that would be the data related to the admission in the hospital or system. These are the ones related to the diagnosis, tests, etc., carried out in that admission or passage through the hospital. In Table 5 we can see some of the characteristics and the origin of the data, whether it has been data generated by processing, whether it has been data extracted or processed by the enrichment phase or whether it is as it appeared in the raw data.

As can be seen, most of these variables are not fully modelled and interpreted. This occurs because, on the one hand, we have many numerical variables whose nature is difficult to represent in the algorithms to be used (association rules). For example, when discretizing these variables to create rules, we must create intervals that do not represent the nature of the variable, so we lose part of the knowledge that we could extract from the data set. To solve the handling of these types of variables, we are going to use the algorithm described in Section 3.3.3, in which, depending on the variables, we can apply labels automatically, using expert knowledge or implementing a function that generates the distribution of the degree of membership. With this processing, we can model these variables more correctly, and we can also label them with fuzzy labels so that the results of our algorithms are more interpretative for end-users.

As an example of some of these variables, we can see in Fig. 5 how the patient's age variable has been transformed into fuzzy labels with different degrees of membership. This division has been made with respect to the segmentation presented in [60] and allows us to represent the nature of the variable and its final interpretation better, as can be seen in some of the results obtained.

Another of the fuzzified variables is the length of time spent in the hospital. This is expressed as the number of days a patient is admitted to a hospital ward. In this, we have used seven linguistic

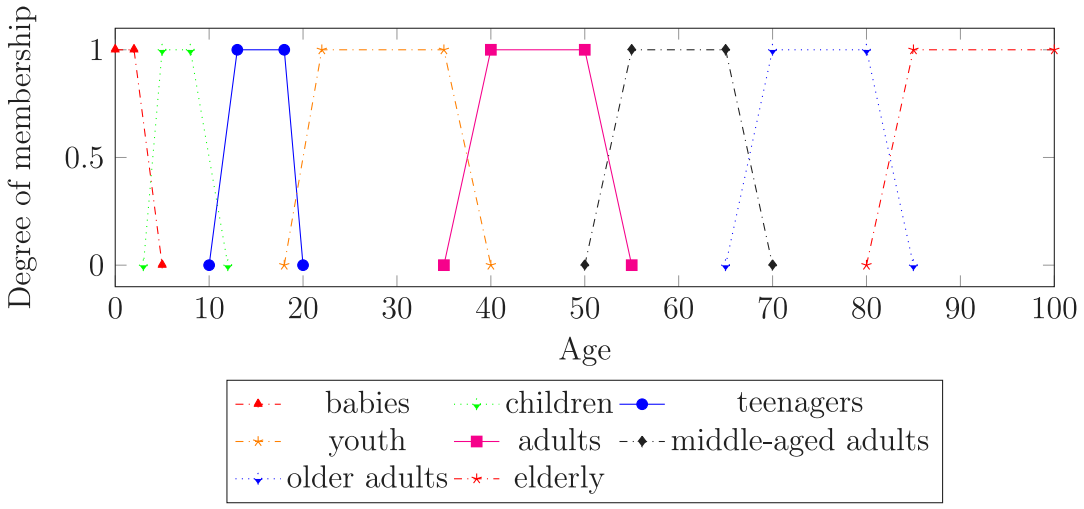

Fig. 5. Example of fuzzification of the *Age* feature.

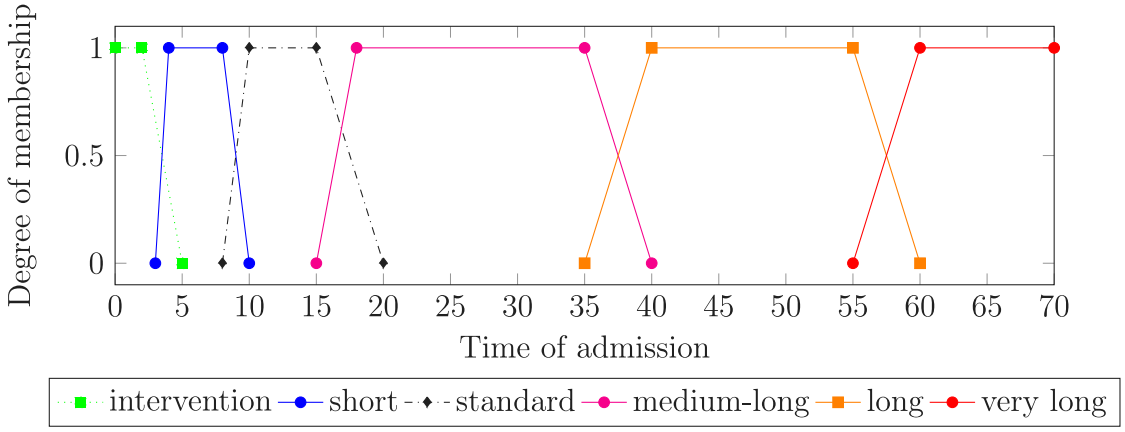

Fig. 6. Distribution of the fuzzy labels as a function of the variable *time of admission*.

labels to express what the stay is like; from the label intervention related to stays of 1 to 3 days, mostly after a simple surgical intervention, childbirth, etc., to very long stays which can last more than two and a half months. These labels can be seen in Fig. 6. If the patient is admitted to the ICU, the variable that stores this information is different, although it has also been fuzzified with the values shown in Fig. 4. The information stored in other variables enables us to extract more knowledge about the database. For example, from diagnoses such as drug dependence, we can infer that the person suffers from an addiction; from problems of nicotine dependence or smoking that the person is a smoker and from other diagnoses, we can infer obesity, alcoholism, etc.

On the other hand, we can obtain levels of complexity or the severity of the person. For example, to infer the level of complexity of a patient's diagnosis, this can be obtained according to the number of diagnoses carried out on admission. A patient may have only one primary diagnosis or up to 20. Of these, we have several kinds, diagnoses obtained on admission (new diagnoses obtained because of their symptoms) or diagnoses that the patient already had. Thanks to this information, we can conclude that the number of diagnoses obtained when a patient is admitted is determined by the number of new diagnoses obtained in that admission. In this way, we have defined their fuzzy labels, as can be seen in Fig. 7. In addition to these examples, we have fuzzified more variables such as the severity of the patient according to the time of admission; we have also processed the ICU times, if the patient is frequently admitted (who has frequent monthly admissions), the weight of the newborns, etc.

4.3. Efficiency analysis and comparison with the sequential approach

As previously mentioned, the proposal was implemented using Spark which enables MapReduce implementation in large data sets. We have carried out different tests in order to be able to analyse the improvement obtained by processing and fuzzifying these data using this framework. The experimental evaluation was carried out on a server cluster with 3 nodes with a total of 102 cores and 320 GB of RAM, running on an operating system with Ubuntu 20. The Spark version was 2.3 using a fully distributed mode with Ambari.

An analysis of the data fuzzification process has been carried out. Depending on the number of processors, different percentages of improvement can be achieved (regarding the computation time see Fig. 8). With the aim of analysing the *speed up* and the *efficiency* [61–63] according to the number of cores, we have used the known measure of speed up defined as [63,64]

$$S_n = T_1/T_n \tag{1}$$

where T_1 is the time of the sequential algorithm and T_n is the execution time of the parallel algorithm using several cores. The efficiency [61–63] can be defined in a similar way as

$$E_n = S_n/n = T_1/(n \cdot T_n) \tag{2}$$

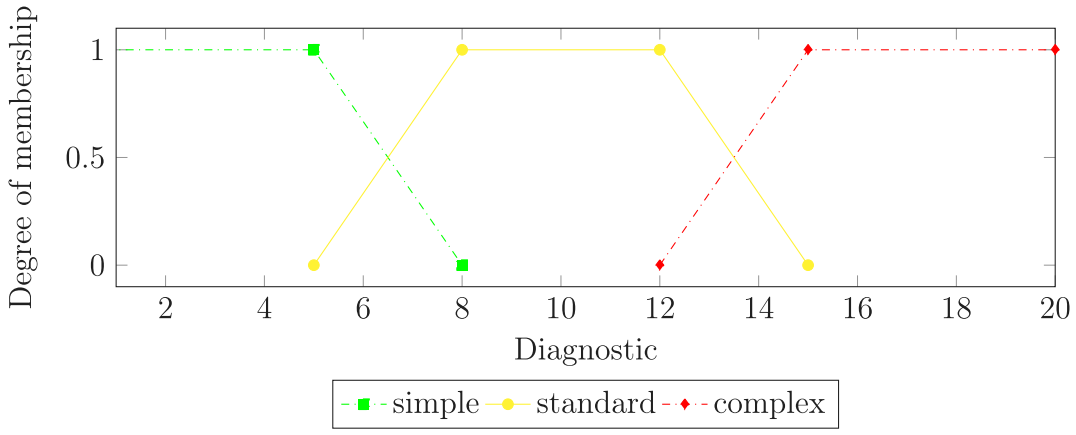

Fig. 7. Distribution of the fuzzy labels as a function of the variable *diagnostic*.

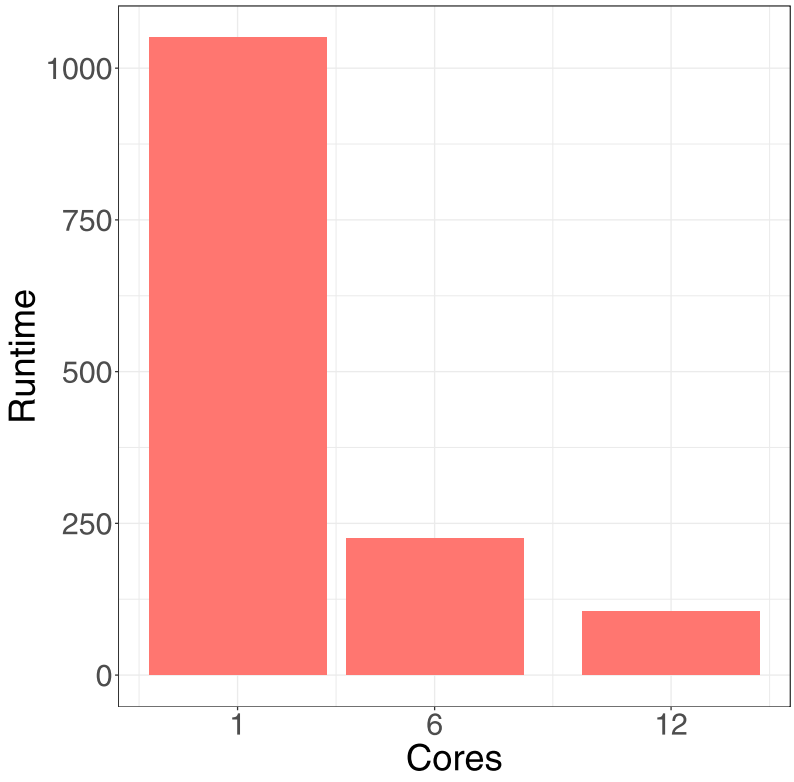

Fig. 8. Speed up of fuzzification process versus number of processing cores.

Figs. 9 and 10 show that the efficiency and speed up are improved as the numbers of cores increases, even when are not optimal. The decrease in efficiency is due to the cores workloads and the network congestion used for the communication amongst the cores.

Using more computational capacity does not represent an improvement from 12 cores because Spark does not divide the data set further due to its size. This implies that if we had more data this process would be able to maintain its efficiency.

In addition, Fig. 9 shows the speedup and evolution of the execution times consumed by the proposal. In this figure, the greatest reduction in calculation time is achieved when the number of processors is 12 is clearly observed.

On the other hand, the use of the fuzzy association rules algorithm has been found to improve the sequential version significantly. In this case, due to the complexity of the process, the algorithm does use the full capacity of the cluster (102 cores).

It can be seen in Fig. 11 how the sequential algorithm of association rules has a worse efficiency and much higher execution times for the two datasets and the experiments with both datasets together. Furthermore, in Fig. 12 we can also see how the behaviour of the algorithm improves from 1 core (sequential) to using the distributed algorithm with more resources (from 24 to 102 cores).

This distributed processing using Big Data improves the processing and computational capacity of large massive data sets. Improving some state-of-the-art proposals such as [47,65], it can be applied to large datasets from hospitals, clinics etc. or even the

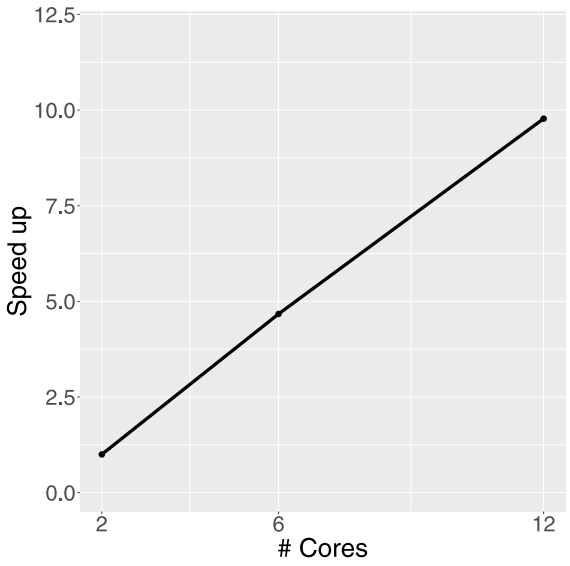

Fig. 9. Time in seconds with different core configurations of the fuzzification algorithm.

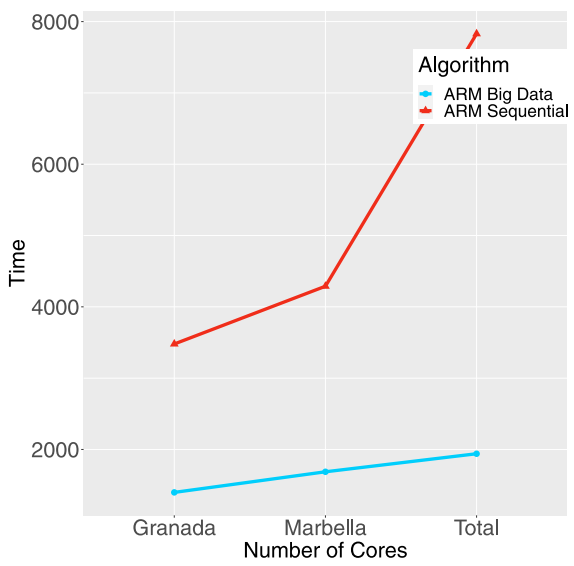

Fig. 11. Execution time of the association rule extraction algorithm with data from each hospital and with all compared to the sequential version (1 core) and in Big Data.

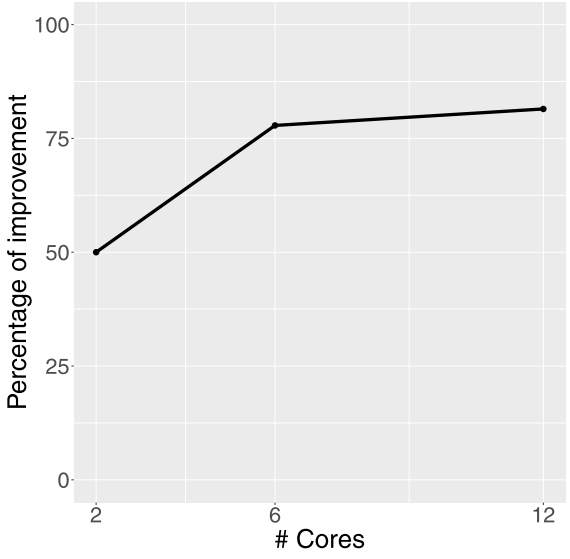

Fig. 10. Efficiency of fuzzification process versus number of processing cores of the rule mining algorithm.

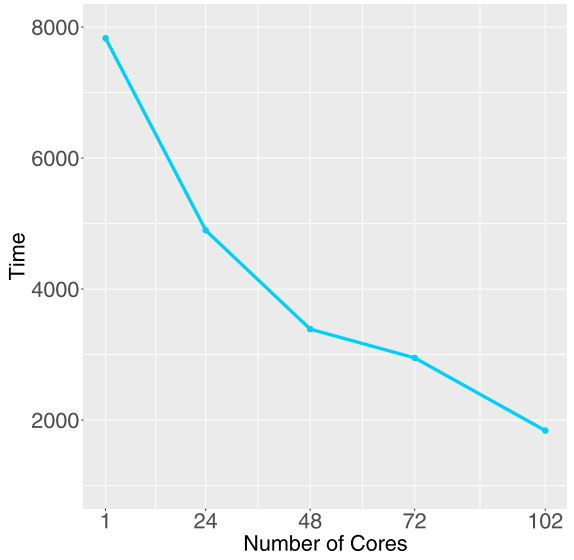

Fig. 12. Efficiency of association rule mining algorithm process versus number of processing cores (1 core→ sequential).

combination of different hospital data sources as has been done in the experimentation.

4.4. Results and discussion

Association rule algorithms have been applied to this processed and enriched data. These algorithms are implemented in Spark to be able to process large data sets because, as we have seen, we have a large number of records with a very high number of features.

The experiments have been applied for different threshold configurations. In particular, we show here the results obtained when the minimum support was set to 0.1 and the minimum confidence to 0.8. We also considered a set of ten equidistributed α-cuts.

The support and confidence thresholds have been set higher than usual due to the high number of resulting rules obtained for lower values. In Fig. 4, the relationship between the amount of support and the number of rules obtained is shown. As can be seen, the number of rules increases as the support increases (see Fig. 13).

The obtained set of rules has enabled us to discover hidden patterns in the relationship between the different diagnoses that appear in the patients and relate these to the patient's characteristics. This type of relationship allows us to study the co-morbidity of the data contained in the system.

Having a look at the discovered patterns, we can highlight different rules. For example:

$$\{Age = long, Time\ of\ admission = medium\_long\} \rightarrow \{Reason\_medical\_discharge = death, Patientseverity = high\} \quad (3)$$

This rule shows how the relationship between elderly patients with a medium to long admission times and that these patients

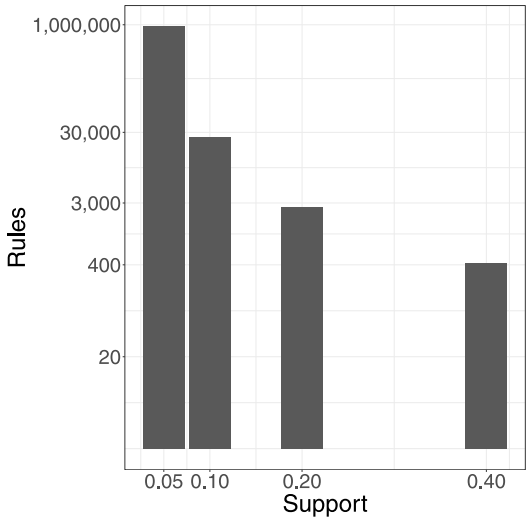

Fig. 13. Number of rules obtained with different parameters of the extraction algorithm.

are seriously ill or die is described. Therefore, we could determine that age and medium-long stays have a grave prognosis.

On the other hand, some rules have been selected, where we can see different behaviour depending on the hospital. For example, this rule obtained in Granada:

$$\{Age = Adult, season = winter, D1 = S83\} \rightarrow \{Diagnostic = simple, Time\_Admission = intervention\} \quad (4)$$

where S83 is equivalent to *dislocation and sprain of joints and ligaments of knee*. This rule provides important information about the relationship between winter knee traumas in Granada; probably because winter sports such as snowboarding and skiing are practised, there are many knee injuries [66]. Moreover, thanks to the diffuse labels we can see how the time of admission of this type of lesions is very short and simple to diagnose because they are not usually associated with other derived diagnoses.

On the other hand, by analysing these resulting rules, we can study the co-morbidity of the different diagnoses. The following rules are obtained by unifying all the data from the two hospitals. In this case, we can see how some diseases such as diabetes generate more complex diagnoses or longer stays. In the first, we can see how the diagnosis of diabetes with cardiac or respiratory problems implies a complex diagnosis and severe patients.

$$\{diabetes, Diseasesofthecirculatorysystem\} \rightarrow \{Diagnostic = complex, Time\_Admission = standard\}$$

The following association rule shows how the diagnosis of alcoholism and drug dependence is related to digestive problems and frequently admitted patients.

$$\{Drugdependence, Alcoholism\} \rightarrow \{Diseasesofthedigestivesystem, frequency = high\}$$

In addition, we have studied, in particular, the COVID19 pandemic-related diagnoses since we have 2020 data from both hospitals in the records of our set. We can find the following rule with a confidence of 0.83, which allows us to relate patients with respiratory diseases who had COVID, needed a ventilator.

$$\{COVID19, Diseasesoftherespiratorysystem\} \rightarrow \{Time\_Admission = long, Time\_Admission = verylong, Dependenceonrespirator, severity = high\}$$

Specifically, this type of patient is admitted for a long stay and has a complex prognosis due to the fact that when assisted ventilation is necessary, the after-effects and recovery are very complex.

5. Conclusions

The discovery and exploitation of information collected from hospitals have attracted attention due to their economic and health impact in the last decade. Big Data offers a suitable framework for the efficient implementation of analysis techniques capable of handling large amounts of data, especially those produced in healthcare systems. In addition, the use of fuzzy logic can improve the interpretability of collected data, offering improved results and interpretation to end-users.

This study has been aimed at the extraction of hidden knowledge from medical records and its analysis and interpretation. For such a purpose, we have implemented a data mining system using the Big Data framework, and we have applied it to different data sets collected from two hospitals in the south of Spain. In particular, we have enriched some features and applied a fuzzification algorithm to improve the performance of data mining techniques, such as association rules. The whole system has been deployed using the Spark platform to analyse such an amount of data generated by the different systems in the hospitals.

The experimentation was conducted with real data from two hospitals to demonstrate the feasibility of our proposal. Our results show the capability of our proposal, discovering interesting rules such as "*alcoholism* and *drug-dependence* is related to *digestive-problems* and *frequent-patients*", and "*the diagnosis of diabetes with cardiac* or *respiratory-problems* implies a *complex-diagnosis* and *severe-patients*", which can be used by end-users to predict and prevent possible diseases, discover relevant relationships between different features and analyse co-morbidity.

The presented system presents significant advantages for the process of analysing medical records in hospitals. This is possible thanks to the distributed computing capability and the innovative data processing process using expert knowledge. This can enable the system to use the data of a hospital or a set of hospitals to process and extract the data contained in the medical records of their patients. It has limitations in that it is adapted to the Spanish standard medical records system, as well as using expert knowledge from Spanish data sources due to the use of this standard.

This is why this research constitutes a starting point, opening up new research lines for the future. The following step consists of using external sources of knowledge such as SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms). This integration will considerably improve the enrichment process of the feature. Another future enhancement concerns visualizing the results which should be more informative and complete for end-users.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research reported in this paper was partially supported by the BIGDATAMED project, which has received funding from the Andalusian Government (Junta de Andalucía) under grant agreement No P18-RT-1765. In addition, this work has been partially supported by the Ministry of Universities through the EU-funded margarita salas programme NextGenerationEU.

References

[1] A.R. Feinstein, The pre-therapeutic classification of co-morbidity in chronic disease, J. Chronic Dis. 23 (7) (1970) 455–468.
[2] P. Fraccaro, D. O'Sullivan, P. Plastiras, H. O'Sullivan, C. Dentone, A. Di Biagio, P. Weller, Behind the screens: Clinical decision support methodologies – A review, Health Policy Technol. 4 (1) (2015) 29–38.
[3] J. Chen, W. Wei, C. Guo, L. Tang, L. Sun, Textual analysis and visualization of research trends in data mining for electronic health records, Health Policy Technol. 6 (4) (2017) 389–400.
[4] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.
[5] B. Kim, R. Khanna, O.O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, Adv. Neural Inf. Process. Syst. 29 (2016).
[6] A. Subasi, Chapter 3 - machine learning techniques, in: A. Subasi (Ed.), Practical Machine Learning for Data Analysis using Python, Academic Press, 2020, pp. 91–202.
[7] L. Zadeh, Fuzzy sets, Inf. Control 8 (1965) 338–353.
[8] C. Fernandez-Basso, M.D. Ruiz, M.J. Martin-Bautista, Extraction of association rules using big data technologies, Int. J. Des. Nature Ecodyn. 11 (3) (2016) 178–185.
[9] P.Y. Taşer, K.U. Birant, D. Birant, Multitask-based association rule mining, Turk. J. Electr. Eng. Comput. Sci. 28 (2) (2020) 933–955.
[10] H. Li, Y. Wang, D. Zhang, M. Zhang, E.Y. Chang, Pfp: parallel fp-growth for query recommendation, in: Proceedings of the 2008 ACM Conference on Recommender Systems, 2008, pp. 107–114.
[11] K. Koperski, J. Han, Discovery of spatial association rules in geographic information databases, in: International Symposium on Spatial Databases, Springer, 1995, pp. 47–66.
[12] D.d. Castro Rodrigues, V. Siqueira, F. Tavares, M. Lima, F. Oliveira, L. Osco, W. Junior, R. Costa, R. Barbosa, Discovering associative patterns in healthcare data, in: Proceedings of Sixth International Congress on Information and Communication Technology, Springer, 2022, pp. 371–379.
[13] C. Ordonez, N. Ezquerra, C.A. Santana, Constraining and summarizing association rules in medical data, Knowl. Inf. Syst. 9 (3) (2006) 1–2.
[14] I.R. Mewes, H. Jenzer, F. Einsele, A study about discovery of critical food consumption patterns linked with lifestyle diseases for swiss population using data mining methods., in: HEALTHINF, 2021, pp. 30–38.
[15] M. Delgado, N. Marín, D. Sánchez, M. Vila, Fuzzy association rules: General model and applications, IEEE Trans. Fuzzy Syst. 11 (2) (2003) 214–225.
[16] Z. Yu, B.C. Fung, F. Haghighat, Extracting knowledge from building-related data. a data mining framework, in: Building Simulation, Vol. 6, Springer, 2013, pp. 207–222.
[17] Z.J. Yu, F. Haghighat, B.C. Fung, Advances and challenges in building engineering and data mining applications for energy-efficient communities, Sustainable Cities Soc. 25 (2016) 33–38.
[18] M. Molina-Solana, M. Ros, M.D. Ruiz, J. Gómez-Romero, M. Martin-Bautista, Data science for building energy management: a review, Renew. Sustain. Energy Rev. 70 (2017) 598–609.
[19] C. Fan, F. Xiao, Mining gradual patterns in big building operational data for building energy efficiency enhancement, Energy Procedia 143 (2017) 119–124.
[20] J. Davis, A. Clark, Data preprocessing for anomaly based network intrusion detection: A review, Comput. Security 30 (6–7) (2011) 353–375.
[21] A. Chandra Pandey, D. Singh Rajpoot, M. Saraswat, Twitter sentiment analysis using hybrid cuckoo search method, Inf. Process. Manage. 53 (4) (2017) 764–779.
[22] J. Jeon, C. Lee, Y. Park, How to use patent information to search potential technology partners in open innovation, J. Intellect. Property Rights 16 (5) (2011) 385–393.
[23] H. Amirkolaee, H. Arefi, Height estimation from single aerial images using a deep convolutional encoder-decoder network, ISPRS J. Photogramm. Remote Sens. 149 (2019) 50–66.
[24] S. Zhang, Q. Shen, C. Nie, Y. Huang, J. Wang, Q. Hu, X. Ding, Y. Zhou, Y. Chen, Hyperspectral inversion of heavy metal content in reclaimed soil from a mining wasteland based on different spectral transformation and modeling methods, Spectrochim. Acta - Part A: Mol. Biomol. Spectrosc. 211 (2019) 393–400.
[25] L. Sun, X. Zhang, J. Xu, S. Zhang, An attribute reduction method using neighborhood entropy measures in neighborhood rough sets, Entropy 21 (2) (2019).
[26] L. Zhang, H. Sun, Z. Rao, H. Ji, Hyperspectral imaging technology combined with deep forest model to identify frost-damaged rice seeds, Spectrochim. Acta - Part A: Mol. Biomol. Spectrosc. 229 (2020).
[27] A. Beam, I. Kohane, Big data and machine learning in health care, JAMA - J. Amer. Med. Assoc. 319 (13) (2018) 1317–1318.
[28] C. Lee, H.-J. Yoon, Medical big data: Promise and challenges, Kidney Res. Clin. Pract. 36 (1) (2017) 3–11.
[29] M. Sabzevari, E. Imani, Separation of movement direction concepts based on independent component analysis algorithm, linear discriminant analysis, deep belief network, artificial and fuzzy neural networks, Biomed. Signal Process. Control 62 (2020).
[30] M. Khan, I. Lali, A. Rehman, M. Ishaq, M. Sharif, T. Saba, S. Zahoor, T. Akram, Brain tumor detection and classification: A framework of marker-based watershed algorithm and multilevel priority features selection, Microsc. Res. Tech. 82 (6) (2019) 909–922.
[31] S. Vajda, A. Karargyris, S. Jaeger, K. Santosh, S. Candemir, Z. Xue, S. Antani, G. Thoma, Feature selection for automatic tuberculosis screening in frontal chest radiographs, J. Med. Syst. 42 (8) (2018).
[32] C. Sakar, G. Serbes, A. Gunduz, H. Tunc, H. Nizam, B. Sakar, M. Tutuncu, T. Aydin, M. Isenkul, H. Apaydin, A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable Q-factor wavelet transform, Appl. Soft Comput. 74 (2019) 255–263.
[33] J. Chen, B. Hu, P. Moore, X. Zhang, X. Ma, Electroencephalogram-based emotion assessment system using ontology and data mining techniques, Appl. Soft Comput. 30 (2015) 663–674.
[34] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, G. Wang, Data processing and text mining technologies on electronic medical records: A review, J. Healthcare Eng. 2018 (2018).
[35] S. Maitra, N. Akter, A. Zahan Mithila, T. Hossain, M. Shafiul Alam, Apriori-backed fuzzy unification and statistical inference in feature reduction: An application in prognosis of autism in toddlers, Adv. Intell. Syst. Comput. 1299 AISC (2021) 233–254.
[36] K. Majumdar, Human scalp EEG processing: Various soft computing approaches, Appl. Soft Comput. 11 (8) (2011) 4433–4447.
[37] S. Lakshmanaprabu, S. Mohanty, S. S., S. Krishnamoorthy, J. Uthayakumar, K. Shankar, Online clinical decision support system using optimal deep neural networks, Appl. Soft Comput. 81 (2019).
[38] A. Nikfarjam, A. Sarker, k. O'Connor, R. Ginn, G. Gonzalez, Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features, J. Amer. Med. Inform. Assoc. 22 (3) (2015) 671–681.
[39] R. Bauder, T. Khoshgoftaar, N. Seliya, A survey on the state of healthcare upcoding fraud analysis and detection, Health Services Outcomes Res. Methodol. 17 (1) (2017) 31–55.
[40] L. Silva, A. Santos, R. Bravo, A. Silva, D. Muchaluat-Saade, A. Conci, Hybrid analysis for indicating patients with breast cancer using temperature time series, Comput. Methods Programs Biomed. 130 (2016) 142–153.
[41] S. Chakraborty, K. Mali, Fuzzy electromagnetism optimization (FEMO) and its application in biomedical image segmentation, Appl. Soft Comput. 97 (2020).
[42] F. Gargiulo, S. Silvestri, M. Ciampi, G. De Pietro, Deep neural network for hierarchical extreme multi-label text classification, Appl. Soft Comput. 79 (2019) 125–138.
[43] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, M. Esposito, Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set, Appl. Soft Comput. 97 (2020).
[44] A. Ponmalar, V. Dhanakoti, An intrusion detection approach using ensemble support vector machine based chaos game optimization algorithm in big data platform, Appl. Soft Comput. 116 (2022).
[45] W. Ai, K. Li, K. Li, An effective hot topic detection method for microblog on spark, Appl. Soft Comput. 70 (2018) 1010–1023.
[46] S. Malla, A. P.J.A., COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets, Appl. Soft Comput. 107 (2021).
[47] M. Delgado, D. Sánchez, M.-A. Vila, Acquisition of fuzzy association rules from medical data, in: S. Barro, R. Marín (Eds.), Fuzzy Logic in Medicine, Physica-Verlag HD, Heidelberg, 2002, pp. 286–310.
[48] F. Ali, S. Islam, D. Kwak, P. Khan, N. Ullah, S.-J. Yoo, K. Kwak, Type-2 fuzzy ontology–aided recommendation systems for IoT–based healthcare, Comput. Commun. 119 (2018) 138–155.
[49] F. Goodarzian, H. Hosseini-Nasab, J. Muñuzuri, M.-B. Fakhrzad, A multi-objective pharmaceutical supply chain network based on a robust fuzzy model: A comparison of meta-heuristics, Appl. Soft Comput. 92 (2020).
[50] C. Fernandez-Basso, M.D. Ruiz, M.J. Martin-Bautista, A fuzzy mining approach for energy efficiency in a big data framework, IEEE Trans. Fuzzy Syst. 28 (11) (2020) 2747–2758.
[51] B. Malysiak-Mrozek, M. Stabla, D. Mrozek, Soft and declarative fishing of information in big data lake, IEEE Trans. Fuzzy Syst. 26 (5) (2018) 2732–2747.
[52] J. Kooij, Adult ADHD: Diagnostic Assessment and Treatment, 2014, pp. 1–292,
[53] G. Mahani, M.-R. Pajoohan, Predicting lab values for gastrointestinal bleeding patients in the intensive care unit: A comparative study on the impact of comorbidities and medications, Artif. Intell. Med. 94 (2019) 79–87.
[54] E. Saleh, J. Błaszczyński, A. Moreno, A. Valls, P. Romero-Aroca, S. de la Riva-Fernández, R. Słowiński, Learning ensemble classifiers for diabetic retinopathy assessment, Artif. Intell. Med. 85 (2018) 50–63.

[55] J. Yoon, D. Jeong, C.-H. Kang, S. Lee, Forensic investigation framework for the document store NoSQL DBMS: Mongodb as a case study, Digital Invest. 17 (2016) 53–65.
[56] M.D. Ruiz, J. Gomez-Romero, C. Fernandez-Basso, M.J. Martin-Bautista, Big data architecture for building energy managament systems, IEEE Trans. Ind. Inf. (2021).
[57] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M.J. Franklin, S. Shenker, I. Stoica, Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing, in: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, USENIX Association, 2012, p. 2.
[58] C. Fernandez-Basso, M.D. Ruiz, M.J. Martin-Bautista, Spark solutions for discovering fuzzy association rules in Big Data, Internat. J. Approx. Reason. 137 (2021) 94–112.
[59] C. Fernandez-Bassso, M.D. Ruiz, M.J. Martin-Bautista, Fuzzy association rules mining using spark, in: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, 2018, pp. 15–25.
[60] S.I. Vuik, E. Mayer, A. Darzi, A quantitative evidence base for population health: applying utilization-based cluster analysis to segment a patient population, Popul. Health Metr. 14 (1) (2016) 1–9.
[61] V.P. Kumar, A. Gupta, Analyzing scalability of parallel algorithms and architectures, J. Parallel Distrib. Comput. 22 (3) (1994) 379–391.
[62] A.Y. Grama, A. Gupta, V. Kumar, Isoefficiency: Measuring the scalability of parallel algorithms and architectures, IEEE Parallel Distrib. Technol. Syst. Appl. 1 (3) (1993) 12–21.
[63] C. Barba-González, J. García-Nieto, A. Benítez-Hidalgo, A.J. Nebro, J.F. Aldana-Montes, Scalable inference of gene regulatory networks with the spark distributed computing platform, in: J. Del Ser, E. Osaba, M.N. Bilbao, J.J. Sanchez-Medina, M. Vecchio, X.-S. Yang (Eds.), Intelligent Distributed Computing XII, Springer International Publishing, Cham, 2018, pp. 61–70.
[64] M.D. Ruiz, J. Gomez-Romero, C. Fernandez-Basso, M.J. Martin-Bautista, Big data architecture for building energy managament systems, IEEE Trans. Ind. Inf. (2021).
[65] J. Calero, G. Delgado, M. Sánchez-Marañón, D. Sánchez, M.A.V. Miranda, J. Serrano, An experience in management of imprecise soil databases by means of fuzzy association rules and fuzzy approximate dependencies, in: ICEIS 2004, Proceedings of the 6th International Conference on Enterprise Information Systems, Porto, Portugal, April 14-17, 2004, 2004, pp. 138–146.
[66] M.J. Jordan, P. Aagaard, W. Herzog, Anterior cruciate ligament injury/reinjury in alpine ski racing: a narrative review, Open Access J. Sports Med. 8 (2017) 71.