Special Issue Paper

Mathematical Methods in the Applied Sciences

Received 14 July 2016

Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/mma.4338
MOS subject classification: 68T35; 68T27; 68T30

Enhancing the conversational process by using a logical closure operator in phenotypes implications

Fernando Benito-Picazo[*†], **Manuel Enciso**, **Carlos Rossi** and **Antonio Guevara**

Communicated by J. Vigo-Aguiar

In this paper, we present a novel strategy to face the problem of dimensionality within datasets involved in conversational and feature selection systems. We base our work on a sound and complete logic along with an efficient attribute closure method to manage implications. All of them together allow us to reduce the overload of information we encounter when dealing with these kind of systems. An experiment carried out over a dataset containing real information comes to expose the benefits of our design. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: conversational systems; feature selection; implications; logic

1. Introduction

A common problem related to knowledge discovering within the clinical context appears when it is necessary to work over datasets with a high number of features (variables or attributes). This situation is known as the curse of dimensionality phenomenon. In these cases, trying to apply data mining techniques in its different approaches (classification, regression, clustering, association rule analysis, etc.) becomes a hard task.

To address this issue, we can find many works in the literature about data reduction, specially in feature selection, that can help us to discard those features not worthy to be considered by means of different criteria. As this respect, several techniques have already been applied such as genetic algorithms, regression, neural networks and many others. All of them guided towards the application of an automated process that is applied at once (batch mode) by feature selection.

The problem of managing large volumes of information is highly present on another hot topic field of knowledge: recommender systems. The major goal of these systems is to help the user when dealing with an extremely high number of alternatives. Recommender systems are present in many different areas of today's society (e-commerce, tourism, films, music, news, etc.) in which large amount of data are pretty much often. Most recommender systems base the retrieval of items in predictions about how suitable is an item to satisfy a user's need. These predictions could be performed from user's preferences, profile, context and so on. To achieve that, different strategies are applied to enclose a recommender system into different types. Best known are content-based, knowledge-based, collaborative filtering and context-based. From the point of view of this work, we centre our efforts not just in the recommendation strategy but also in the process of obtaining a recommendation.

As is usually the case, in order to properly make a selection of items within recommenders, the user needs to introduce information over and over. That used to be something of a chore because many items may be too much detailed drawing the high-dimensionality problem mentioned earlier. An example of this situation appears when medical professionals try to elaborate a diagnosis checking symptoms from a vast list of possible ones.

One solution to tackle this issue in the field of recommender systems is the trend of conversational systems (critiquing recommender systems overall). In these systems, an iterative process is applied in which the user sets one or more features for items to fit. From this input and using different techniques, the system progresses choosing (or even predicting) subsets of items that agree with the user's preferences, until it comes to a final output with a suitable size. The problem is that, if the dataset presents a high dimensionality, the number of steps until we acquire a fitting recommendation set could be huge.

Universidad de Málaga, Andalucía Tech, Málaga, Spain
[*] *Correspondence to: Fernando Benito-Picazo, Universidad de Málaga, Andalucía Tech, Málaga, Spain.*
[†] *E-mail: fbenito@lcc.uma.es*

The solution we proposed in this work is to manage the problem of the high dimensionality by means of a feature selection process guided by the user (human expert) within a conversational system. To achieve that, we base the approach on a novel management of implications and closure implementation. Our proposal avoids the problem mentioned in [1] remarking that batch mode feature selection systems extract them randomly. Their only criterion in the selection is the feature predictive capacity, without considering the medical knowledge. For instance, the Computer Feature Selection over Cleveland dataset discards some key features such as cholesterol, age or ECG characteristics. On the contrary, we surpass this issue by making an expert-driven automated selection.

Our major goal is to enhance the diagnosis process preserving the accuracy of the results (that is granted when using implications) and accelerating the process by reducing the necessary steps in the dialogue. Besides, the reduction of the complexity within the process is also a must, so we can obtain results timely and in due form.

As already mentioned, this work is based on implication management. Nonetheless, it is necessary to clarify that it is out of the scope of this work to describe how to extract these implications, because there are already data mining techniques specialised on this task [2].

The set of implications is the heart of the knowledge in the process we propose. We support our approach on a sound and complete logic named simplification logic (SL_{FD}), presented in several previous works, which was designed to develop deduction methods. In particular, we propose to use SL_{FD} attribute closure algorithm as the core of a feature selection framework. This algorithm is used as a basis for reasoning over the features (symptoms, phenotypes, signs) of the items (diseases) of the dataset we are going to use, reducing the number of steps in the conversation to reach a suitable diagnosis.

The key point of our framework comes from the SL_{FD} closure nature that renders a set of features that corresponds with the closure and, in addition, a new implication set corresponding with the knowledge not already used in the conversation (selection o diagnosis) process. Such implication set can be obtained by using other methods but, with SL_{FD} closure method, we compute it in linear time. This low complexity along with the reduction of the number of steps provide an overall time reduction for the process.

Our framework has been tested on a dataset that puts together real information about diseases and phenotypes. Particularly, we have selected a set of haematological diseases with phenotypes related to them. In the experiment, several metrics to evaluate the system's performance have been applied, and results have come to demonstrate the highlights of our framework.

The remainder of the paper is organised as follows: Section 2 draws attention about the state-of-the-art references and their motivations along with a brief introduction of the main elements of our approach. Section 3 brings us a detailed explanation about how the information will be managed, the notions of the SL_{FD} that apply in this approach and the specification and the usefulness of the closure algorithm. The conversational processes our approach carries out along with major benefits it reaches are presented in Section 4. Section 5 presents an empirical test considering a real case in a medical environment and the metrics used to evaluate the performance. Conclusions and future works close the paper.

2. Related works

Our approach tries to solve the problem of searching information in high-dimensionality datasets by combining techniques from diverse fields (recommender systems, feature selection, logic) in a knowledge-based framework. In this section, we analyse some outstanding related works in each area.

As mentioned, in general terms, the problem we aim to solve is to search a precise result (e.g. a diagnosis) from a high volume of data (e.g. a diseases dataset) without previous user information. By these means, our approach is inspired in one of the solutions proposed in this field, the so-called conversational recommender systems [3]. These are closely related with the concepts of critiquing recommender systems [4] and information recommendation [5]. In these systems, recommendation is generated by means of a dialogue with the user that allows an incremental elicitation of preferred item features, that is, user requirements are directly elicited within a recommendation session. An interesting work in this area is [6], which states the suitability of knowledge-based approaches for conversational processes. In particular, these authors use constraint-based reasoning, instead of our logic-based approach. Besides, this work deals with the concept of query tightening, analogous to the one applied in our proposal.

Another remarkable work is [5], which shares our aim to decrease the conversation length (number of steps). The theoretical framework they use is a model of attribute dominance with two versions, a qualitative and a quantitative one. These authors propose metrics about conversation number of steps and pruning rates, both of them very similar to the ones used in our work. Nevertheless, they present an experiment focused on tourism.

Regarding critiquing recommender systems, Chen and Pu [7] explain how user self-motivated attribute selection enable users to achieve a higher level of decision accuracy than system-proposed attribute selection. This fact supports our approach in which human expert guides the conversation and the feature selection process.

The main elements of our theoretical framework are SL_{FD} as well as the concept of attribute closure and the algorithm we have defined to compute it. One of the main advantages of this algorithm is the reduction of the time required by the selection process. For a better understanding, works related with the theoretical basis are exposed in the next section.

The starting point of our selection process is the set of implications (also named exact association rules) derived from the working dataset. To acquuire these implications, association rule mining is needed [8]. It is easy to find in literature works evaluating association rule mining techniques both in the biological–medical [9] and recommender systems [10] fields.

Once association rules had been inferred from the dataset, they can be used in the main medical tasks, such as screening, diagnosis, treatment, prognosis, monitoring and management [11, 12]. For example, Nahar *et al.* [13] uses association rule mining to analyse factors related with heart diseases. In the same way, many other works use association rules to build a decision support system or prediction model [14–17].

Closer to our work, Mansing *et al.* [18] deals with the fact that the number of association rules used to be very high. This work proposes an association rule pruning mechanism, based on well-known concepts as support, confidence and reliability. A similar idea is proposed in [19] and [20], although the latter applies temporal abstraction in the decision support system. Our approach includes an association rule filtering process based on the closure concept. Hu *et al.* [21] proposes the use of association rules for predicting combination of alarms generated by bedside monitors. These authors identify frequent alarm combinations and then carry out a variance analysis to measure the data mining process performance. Regarding our work, it should be noticed that they use a closure-like concept to define a heuristic that controls combination size. In any case, we must remark that our work is not strictly comparable with the previously cited ones, because we do not aim to build a prediction model but to improve a conversational process.

In the introduction, we affirm that our approach may be described as an expert-driven feature selection process. As Fang *et al.* [22] exposes, feature selection is a problem profusely studied in the literature. So, there are a number of solutions based on statistical techniques such as principal component analysis, linear discriminant analysis and independent component analysis, all of them suitable to deal with high-dimensionality datasets. Besides, other authors use evolutionary computing techniques such as particle swarm optimisation [23, 24]. Nevertheless, it should be remarked that most of feature selection works are focused on batch (non-interactive) processes. This way, the system, without user intervention, determines and selects more relevant features according to their predictive significance.

Recently, some papers dealing with iterative feature selection processes have been published. For example, Fialho *et al.* [25] proposes a tree feature selection process based on fuzzy modelling (and fuzzy rules). They use two tree searching techniques (sequential forward selection and sequential backward elimination). This interesting work achieves good results in well-known metrics such as Area Under Curve (AUC) or accuracy, but requires many inputs when the dataset has a high number of features. Analogously, Shilaskar and Ghatol [26] apply the same techniques but using Support Vector Machine (SVM) as classifier.

Some authors analyse the relation between features as means to reduce the dataset, for example, using correlation feature selection as a basis. This is the case of [27], although they propose a batch, non-interactive, process.

Closer to an interactive feature selection, Li *et al.* [28] define a selection process in which the input is a feature stream. They use information theory techniques (mutual information, conditional mutual information, entropy) to build a prediction model and achieve good results about the number of selected features. Although they process the features sequentially, the work is not oriented to perform as a conversational selection process.

In this way, we must remark [29], which highlights the importance of an online feature selection, so that features are processed one-by-one. This work is aimed to an extremely high-dimensionality context (in the order of millions of features) and is based on information theory. They use pairwise correlation analysis as a mean to remove redundant features. The goal of this proposal is to build a prediction mode with scalability as its main advantage.

In [30], an interesting work is presented that, as ours, deals with association rules and feature selection. Nevertheless, they base the feature extraction on methods such as partial least squares or principal component analysis, and it is not directly comparable with our approach.

3. Knowledge representation and automated deduction

In this section, we address the issue of specifying and efficiently managing the information. Knowledge-based systems have to balance both commitments to obtain, at the same time, powerful and efficient techniques. In our opinion, one of the tools that has shown a better behaviour is implications. They combine a very simple and natural way to write if-then-rules with an efficient and automated management. One evidence supporting our choice is its widespread use in different areas: databases, logic programming, formal concept analysis, artificial intelligence and so on.

In this work, knowledge is stored by considering implications, following the interpretation adopted in formal concept analysis [31], because of their simplicity. In this area, implications are inferred from datasets that are considered as binary relations between a finite set of objects and their attributes (depicted by rows and columns, respectively). Such interpretation is the following: Given a formal context K, an implication is an expression $A \rightarrow B$, where A and B are subsets of attributes, and it is said to be valid in K if and only if every object that has all the attributes from A also has all the attributes from B. Example 1 illustrates the knowledge captured by using implications.

Example 1

Let K be the formal context described in Table I showing the 22 common viruses and their usual ways of transmission.

The implications that hold in this dataset are as follows:

$\texttt{Blood} \rightarrow \texttt{Sexual}$
$\texttt{Fluids} \rightarrow \texttt{Vertical}$
$\texttt{Direct,Sexual,Droplet} \rightarrow \texttt{Faecal}$
$\texttt{Faecal,Sexual} \rightarrow \texttt{Direct,Droplet}$
$\texttt{Droplet} \rightarrow \texttt{Direct}$
$\texttt{Respiratory} \rightarrow \texttt{Direct,Droplet}$
$\texttt{Direct,Vertical} \rightarrow \texttt{Sexual}$
$\texttt{Faecal,Direct} \rightarrow \texttt{Droplet}$

In this way, implications allow to express a strong relation between two subsets of attributes of our system. Moreover, such information can be interpreted in a natural way. For instance, the last implication tells us that if a virus is transmitted by a direct contact and with faecal transmission, we also have to be vigilant about the sneezes.

Our proposal to integrate implications into the conversational issue is based on the SL_{FD} [32], which constitutes a sound and complete logic. As we shall see, such a strong basis allows us to include a reasoning method in the dialogue process. As mentioned, we built our framework on implications, which constitutes the main element of SL_{FD} language. It is formally defined as follows:

Table I. Viruses and usual way of transmissions dataset.

	Faecal	Direct	Vertical	Sexual	Respiratory	Saliva	Fluids	Droplet	Blood
Adenovirus	×	×		×				×	
Coxsackievirus	×	×			×			×	
Epstein-Barr						×			
Hepatitis A	×								
Hepatitis B			×	×			×		
Hepatitis C				×					×
Herpes type 1		×				×			
Herpes type 2			×	×					
Cytomegalovirus			×				×		
Herpesvirus type 8				×		×			
HIV			×	×					
Influenza		×						×	×
Measles virus		×						×	
Mumps virus		×						×	
Papillomavirus		×	×	×					
Parainfluenza		×						×	
Poliovirus	×								
Rabies		×						×	
Respiratory syncytial		×						×	
Rubella		×			×			×	
Varicella zoster		×						×	

Definition 1
Let M be a finite set, the formulae of SL_{FD} are expressions, named implications, of the form $X \to Y$, where X and Y are subsets of M.

From now on, we use lower case letters to denote the elements in M while uppercase letters denote its subsets. We use the standard notation and symbols of set theory. For the sake of readability, inside of a formula, $X - Y$ denotes the set difference operator $X \smallsetminus Y$, and XY denotes the union operator $X \cup Y$.

Implications are interpreted in a conjunctive way, that is, they correspond to formulas $a_1 \wedge \ldots \wedge a_n \to b_1 \wedge \ldots \wedge b_m$ where propositions $a_1, \ldots a_n, b_1, \ldots, b_m$ are elements of the set M. The interpretation is the following:

Definition 2
Let O and M be two finite sets, named objects and attributes, respectively, and I a relation in $O \times M$. An implication of SL_{FD} $X \to Y$, where X and Y are subsets of M, is valid in I if and only if

$$\{o \in O \mid (o, x_i) \in I \, \forall x_i \in X\} \subseteq \{o \in O \mid (o, y_j) \in I \, \forall y_j \in Y\}$$

Apart from its natural way to express knowledge as a rule, implications provide a logic reasoning and inference. Their symbolic management was originally proposed in [33]. However, because of the central role that transitivity plays in that axiomatic system, the development of executable methods to solve implications problems has rest on indirect methods. The introduction of the SL_{FD} opened the door to the development of automated reasoning methods directly based on its novel axiomatic system [34, 35]. The axiomatic system of SL_{FD} is introduced as follows:

Definition 3
The axiomatic system of SL_{FD} considers reflexivity as axiom scheme

$$[\mathtt{Ref}] \quad \frac{}{A \to A}$$

together with the following inference rules called fragmentation, composition and simplification, respectively.

$$[\mathtt{Frag}] \frac{A \to BC}{A \to B} \qquad [\mathtt{Comp}] \frac{A \to B,\ C \to D}{AC \to BD} \qquad [\mathtt{Simp}] \frac{A \to B,\ C \to D}{A(C - B) \to D}$$

As we mentioned earlier, in [32], we defined, in the usual way, the semantic entailment ($\Gamma \models A \to B$) and syntactic derivation ($\Gamma \vdash A \to B$). Because we also proved the soundness and completeness of this logic, both notions can be equivalently used. We also introduced the notion of equivalence between sets of implications $\Gamma_1 \equiv \Gamma_2$ iff for all $A \to B \in \Gamma_1$, we have that $\Gamma_2 \vdash A \to B$ and vice versa.

We remark that SL_{FD} language considers as valid formulae those ones where any of their two parts can be the empty set, denoted $A \to \top$ and $\top \to A$. Their meanings were discussed in [32]. In that work, we also introduced the following result where the derivation of an implication $A \to B$ is reduced to the derivation of the formula $\top \to B$ having $\top \to A$. This result will be used later in the design of our novel closure method.

Proposition 3.1
For any $\mathit{\Gamma}$ and for all $X, Y \subseteq M$, $\mathit{\Gamma} \vdash X \to Y$ if and only if $\mathit{\Gamma} \cup \{\top \to X\} \vdash \top \to Y$

The syntactic derivation provides an automated management of implications. In particular, it can be used to solve the so-called implication problem: Given a set of implications Γ and an implication $A \to B$, we want to answer whether $A \to B$ is deduced from Γ. This problem can be approached by using the closure operator.

Definition 4
Let $A \subseteq M$ be a set of attributes and Γ a set of implications; we define its closure with respect to Γ as the maximum subset $A^+_\Gamma \subseteq M$ such that the $\Gamma \vdash A \to A^+_\Gamma$.

Moreover, the set A is named closed iff we have $A^+_\Gamma = A$.

Implication problem has been traditionally tackled by using a basic method that receives $A \subseteq M$ as input and exhaustively uses the subset relation by iteratively traversing Γ and adding new elements to the closure. This method was proposed in the 1970s [36], and it is sketched in Algorithm 1.

Algorithm 1: Standard Closure

```
Data: $\mathit{\Gamma}, A$
Result: $A^+_\Gamma$
begin
    $A^+_\Gamma := A$
    repeat
        $A' := A^+_\Gamma$
        foreach $X \to Y \in \mathit{\Gamma}$ do
            if $X \subseteq A^+_\Gamma$ and $Y \not\subseteq A^+_\Gamma$ then
                $A^+_\Gamma := A^+_\Gamma \cup \{Y\}$
    until $A^+_\Gamma = A'$;
    return $A^+_\Gamma$
```

Later, several authors have developed several methods by using different techniques, efficiently solving this problem in linear time. In [37], the authors show that the complexity of closure problem is $\mathcal{O}(|\mathcal{A}|\,|\mathit{\Gamma}|)$. They also mention that 'in the literature, $\mathcal{O}(|\mathcal{A}|\,|\mathit{\Gamma}|)$ is usually considered as the order of the input. From this point of view, this is a linear time complexity for the computation of the closure of a set of attributes'.

In [38], we presented an attribute closure method closely tied to the SL_{FD} axiomatic system. We also showed that our method has a better performance than those based on classical closure. In this paper, we are going to use this method taking advantage of its novel characteristic.

Apart from having a strong base and good performance, one innovative feature of our method is that its output is twofold: besides the A^+_Γ set constituting the closure of the input attribute set A, it also renders a reduced set of implications that encloses the semantics that is outside the set A^+_Γ. We would like to remark that these two inputs are computed in linear time, because the subset of reduced implications is computed by the algorithm at the same time it computes the attribute closure.

The kernel of this closure method is the existence of three equivalences that can be enunciated by using SL_{FD}:

- Equivalence I: If $U \subseteq W$, then $\{\top \to W, U \to V\} \equiv \{\top \to WV\}$
- Equivalence II: If $V \subseteq W$, then $\{\top \to W, U \to V\} \equiv \{\top \to W\}$
- Equivalence III: If $U \cap W \neq \varnothing$ or $V \cap W \neq \varnothing$, then $\{\top \to W, U \to V\} \equiv \{\top \to W, U - W \to V - W\}$

The closure method works as follows. To compute the attribute closure of A with respect to Γ, the method is triggered by the seed formula $\top \to A$ and, exhaustively executing these three equivalences, it modifies this implication rendering $\top \to A^+_\Gamma$ and, at the same time, a reduced set of implications Γ'. The method is described in Algorithm 2.

We end this section with an illustrative example.

Example 2
Let Γ be the set of implications from Example 1. We show how SL_{FD} closure computes the closure of the attribute `Droplet`. In Table II, we show the application of the equivalences to each implication in the Γ set and how the set of attributes grows. We would like to remark that, in this example, only one repeat loop is needed.

As a final conclusion of this example, our method received the set Γ and `Droplet` attribute and renders as output the pair: $\langle\{\texttt{Droplet, Direct}\}, \{\texttt{Blood} \to \texttt{Sexual}; \texttt{Fluids} \to \texttt{Vertical}; \texttt{Sexual} \to \texttt{Faecal}; \texttt{Vertical} \to \texttt{Sexual}\}\rangle$. That is to say, $\texttt{Droplet}^+$ and the reduced set of implications that stores the knowledge complementing the closure set.

Algorithm 2: The SL_{FD} Closure

```
Data: $\Gamma, X$
Output: $\langle X^{+}, \Gamma' \rangle$
begin
    $\Delta := \langle X, \Gamma \rangle$
    repeat
        $\Gamma' := \Gamma$
        foreach $Y \to Z \in \Gamma$ do
            if $Y \subseteq X$ then      /* Equivalence I */
                $\Delta := \langle XZ, \Gamma \setminus \{Y \to Z\} \rangle$
            else if $Z \subseteq X$ then      /* Equivalence II */
                $\Delta := \langle X, \Gamma \setminus \{Y \to Z\} \rangle$
            else if $Y \cap X \neq \varnothing$ or $Z \cap X \neq \varnothing$ then      /* Equivalence III */
                $\Delta := \langle X, (\Gamma \setminus \{Y \to Z\}) \cup \{(Y - X) \to (Z - X)\} \rangle$
    until $\Gamma' = \Gamma$;
    return $\Delta$
```

Table II. An application example of SL_{FD} closure.

Closure	Implication	New implication	
Droplet			
Droplet	Blood → Sexual	Blood → Sexual	-
Droplet, Direct	Droplet → Direct	×	Equiv. I
Droplet, Direct	Fluids → Vertical	Fluids → Vertical	-
Droplet, Direct	Respiratory → Direct, Droplet	×	Equiv. II
Droplet, Direct	Direct, Sexual, Droplet → Faecal	Sexual → Faecal	Equiv. III
Droplet, Direct	Direct, Vertical → Sexual	Vertical → Sexual	Equiv. III
Droplet, Direct	Faecal, Sexual → Direct, Droplet	×	Equiv. II
Droplet, Direct	Faecal, Direct → Droplet	×	Equiv. III

SL_{FD}, simplification logic.

4. Conversational process

Once we have presented our basis, in this section, we depict a detailed explanation of the dialogue process system along with a basic schema to facilitate the comprehension.

First of all, we depart from the premise that we have a dataset containing diseases and phenotypes and the set of implications that holds on it. This is considered the starting point at which this work begins and, as mentioned before, it stays out of the scope of this work. From there on, the dialogue process will go along through the following points:

(1) Once we count on this information, the user starts interacting with the system by selecting a phenotype she suspects a disease to be related to. In this paper and with the intention of illustration, we are limiting the number of phenotypes to be selected to one at a time in each step of the conversation. This way, we can easily appreciate how the system performs. Yet a generalisation system considering more than one phenotype per step showed a similar behaviour but going faster.
(2) Then, the process flows into the closure algorithm calculating both the phenotype's closure set for the selected one and also the reduced set of implications that corresponds with the complement of this closure.
(3) Once the closure algorithm has finished, a first possible diagnosis is shown. This diagnosis shows the list of diseases in the dataset agreeing with the selected phenotypes. In order to acquire this list, the system launches a query to the database requesting those diseases that verify the selected phenotypes.
(4) At this point, the user can stop the dialogue in the case that she is already satisfied with the result (a list of diseases), or she can go ahead trying to acquire a more reduced diagnosis.
(5) Finally, the user can return to step 2 selecting new phenotypes until she acquires a satisfying set of disease result (possible diagnosis) or the system runs out of phenotypes.

The main contribution of this paper is the benefits of reducing the number of available phenotypes (in general, the attributes space), clarifying in each step the user scene. For further steps in the dialogue, we reduce the number of available phenotypes deleting those included in the closure set. As a consequence, some closure's phenotypes remain hidden to the user, in an intelligent and consistent way, because they are already implicitly included. This saves the user unnecessary efforts about information overload. However, even that this could be accomplished by classic closure algorithms, the major novelty of our method is that, *at the same time*, we also reduce

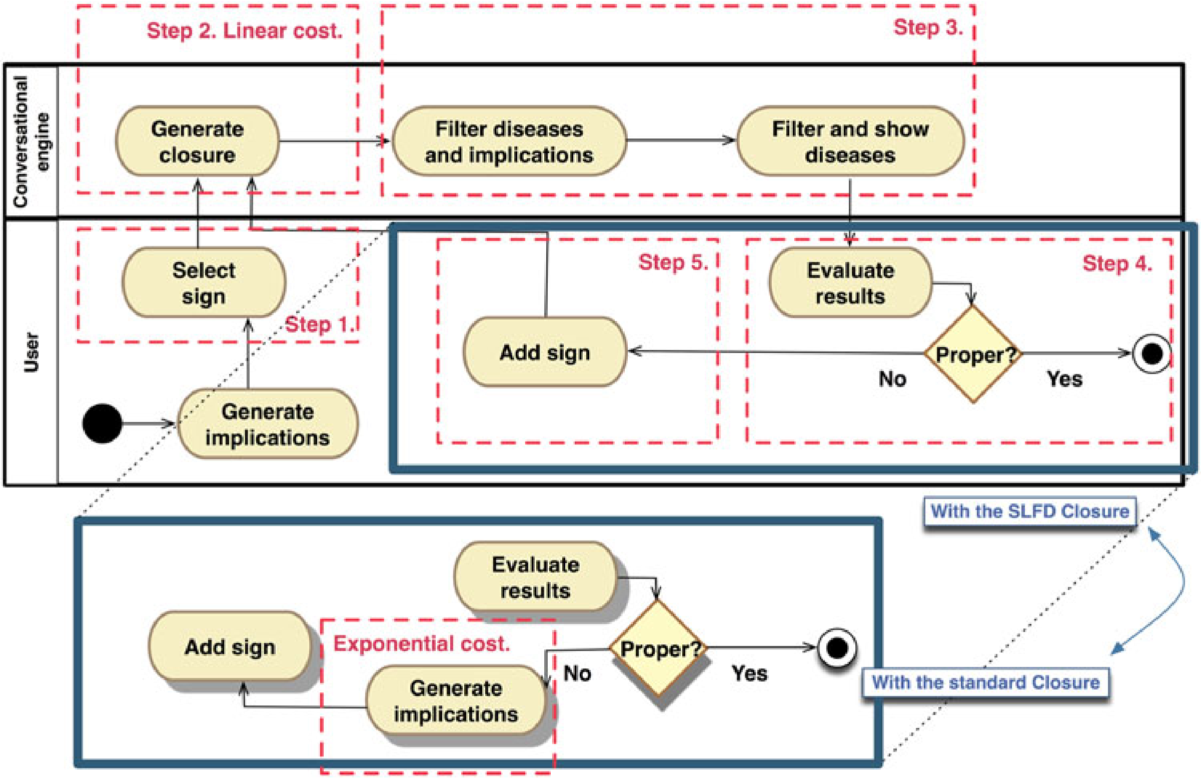

Figure 1. Dialogue process schema using simplification logic (SL_{FD}) closure versus standard closure.

the number of implications. This fact totally places us in a privileged position, overtaking the hard cost of a data mining process to extract the new set of implications for the reduced dataset after each searching step, as necessary when using classical implementations. On the contrary, in every refining attempt in the dialogue using our approach, we do not need to start the process from the beginning but continuing from there, where both symptoms and implications have been decreased. Consequently, the process overcomes the data mining costs preserving a linear complexity along the dialogue, and the interaction becomes truly faster. Figure 1 depicts the aforementioned steps by using the activity diagram of the UML language.

As mentioned earlier, in step 0, we need to count on the inferred set of implications within the dataset (generate implications activity). Although this is an exponential task, it is independent of the closure strategy we are willing to use. Thereafter, we execute step 1 and select the observed sign to continue with the execution of the closure method itself (generate closure activity). Observe that either our closure implementation or classical ones are able to perform in linear time. Now, however, it is remarkable that no matter how many steps we want the dialogue to go further, there is no need of mining any new set of implications; we already have it each time our closure implementation is applied because it is narrowing both attributes and implications at the same time thanks to the output of SL_{FD} closure (filter diseases and implications activity).

Nevertheless, in the case of classical closure implementations, the disadvantage arises as follows. Because only attributes are narrowed each time the closure applies, in order to proceed with the next step of the conversation, we need to generate a new set of implications accordingly to the new narrowed set of attributes. Therefore, this is an exponential task that is mandatory after every step of the dialogue. As a consequence, the overall complexity of the process becomes exponential.

Once the advantages of our method have been exposed, we proceed now with an example that illustrates the achievement of our approach about avoiding the exponential complexity of the conversational process.

Example 3

In the main experiment of this paper (which will be explained in Section 5.2), we have used a real dataset matching information about diseases and phenotypes. From the different runs of this experiment, five steps have been the maximal length the dialogue reached in a simulated conversation. Therefore, we are going to compare here this dialogue with such limit situation in two possible scenarios, the use of SL_{FD} closure and the use of any other classical closure implementations regarding the implication mining task. Here, we do not worry about the execution time of the closure method itself, because it is common to both approaches and not relevant compared with the implication mining cost.

This comparison is depicted in Table III. This table counts on five columns. First one indicates the number of steps along the conversation. Second one denotes the phenotype selected at each step of the dialogue. Columns 3 and 4 are the most important ones; they show the time needed to enumerate the new set of implications after each new step using classical closure implementations (column 3) and SL_{FD} closure (column 4). Last column shows the cardinal of the regenerated set of implications. At length, we show the time needed in the process using either a classical closure implementation or SL_{FD}.

Table III. Total time saved enumerating the set of implications by using SL_{FD} closure.

Step	Phenotype selected	Standard closure	SL_{FD} closure	Number of implications
1	HPO_7	22 s 181 ms	—	8.578
2	HPO_639	21 s 526 ms	—	8.270
3	HPO_2910	20 s 324 ms	—	7.988
4	HPO_1250	19 s 627 ms	—	7.534
5	Dialogue ends	—	—	—
	Total time saved	83 s 658 ms	—	

SL_{FD}, simplification logic; HPO, Human Phenotype Ontology Consortium.

These measures have been obtained by virtue of the application of the specific package for R language named *Arules: Mining Association Rules and Frequent Itemsets*[‡]. Moreover, the hardware configuration used goes as follows: Intel Core 2 Duo 2.6 Ghz, 4 Gb RAM running over Windows 7.

The significant overall reduction of time obtained by using SL_{FD} closure is because of the linear calculation of the new set of implications to be used in the next step, whereas classical implementations suffer from regenerating implications again and again in each step.

5. Application of SL_{FD} conversational method to hematologic diseases selection

In this section, we describe the promising results obtained by our method in a real case. First, we establish the metrics defined to measure the benefits of our method and, later, we describe the selected dataset and the results of the experiment over it.

5.1. Evaluation metrics

When trying to enhance the interaction within a conversational system, evaluating the length of the dialogue for a typical query could be considered the most basic test [39]. Another popular measures to evaluate the effectiveness of recommender systems point to *precision* and *recall* measures [40]. Precision, defined as P = TP/(TP + FP), where TP means the number of true positives, FP the number of false positives and FN the number of false negatives, determines the fraction of relevant items retrieved out of all items retrieved. On the other hand, recall R = TP/(TP + FN), which determines the fraction of relevant items retrieved out of all relevant items. These two popular measures may not shed light on the matter of evaluating this approach, and the reason is twofold. First, every disease belonging to the list of resulting diagnoses agrees with the symptoms selected by the user as the database queries launched to retrieve diseases reflect these constraints. And second, after every loop of the process, we retrieve all the existing diseases in the dataset that hold with the symptoms selected,that is, all the relevant elements. Something similar occurs with other historical measures like mean absolute error or root mean square error. These two measures based on ratings have no place in this approach because no ratings are considered in order to conduct the conversation and the final items retrieval. In addition, there is no need to explicitly consider an accuracy measure because we do not build a prediction model, instead the use of implications ensures full accurate results. Fortunately, there are also others measures that could fairly reflect the statistics of the experiments. We continue describing those ones considered in this work.

5.1.1. Number of steps (N). This metric evaluates the actual length in steps of the conversation. It is interesting in the sense that it offers a rapid overview of the length of the interaction between the user and the conversational system. That gives an idea of whether the conversation has quickly satisfied the user or too many steps were needed. At this regard, within the following experiments, we will consider, without loss of generality, the user satisfied when a result of five (or less) diseases is returned. We set this limit because it deems advisable in order to constitute a proper diagnosis. Simultaneously, it represents the number of attributes (phenotypes) requested by the user because we are hitherto selecting one attribute at a time.

$$N = |\text{Selected attributes}|, \quad \text{where } |A| \text{ represents the cardinal of A.}$$

5.1.2. Speed of pruning at step i (S_i). This metric evaluates the percentage of the attributes the user is saving over the course of the conversation, accumulating from one step to another. When using this metric, we are willing to noticed whether the pruning rates have been better at the first steps of the conversation or at the last ones. Overall, it is a metric to measure how *fast* the system removes the overload of information. Notice that as mentioned before, we are taking one attribute at a time in this approach.

$$S_i = \frac{|\text{Attribute Closure}|_i - i}{|\text{M}|}, \quad i = 1, \ldots, N, \text{ and M represents the whole set of attributes as shown in Section 3.}$$

5.1.3. Attributes pruning (P). This last metric is equal to the speed of pruning but taking values at the end of the dialogue. It represents the percentage of the attributes that have been removed from the original set throughout the conversation. The pruning of these attributes is consequent because they are implicit by the user's selected ones. Formally,

$$P = S_N$$

Once all the basis have been defined and the metrics are explained and formulated, we are ready to begin the experiments.

[‡] *https://cran.r-project.org/web/packages/arules/index.html.*

Table IV. Diseases and symptoms dataset (extract).

Disease ID	HPO_1249	HPO_1250	HPO_1251	HPO_1252	HPO_1254	HPO_1257	…
274000		X					
275630	X		X				
277380				X	X		
300884		X		X			
300322	X			X		X	
…							

HPO, Human Phenotype Ontology Consortium.

5.2. Experiments and results

In this section, we are going to perform experiments on a real-world dataset with a substantial amount of information. First of all, we are going to provide information about the dataset we are going to work with.

The source from which we have extracted the data is the Human Phenotype Ontology Consortium[§] (HPO). As can be read in their web page: *'HPO* [41] *aims to provide a standardized vocabulary of phenotypic abnormalities encountered in human disease. Each term in the HPO describes a phenotypic abnormality. The HPO is currently being developed using the medical literature, Orphanet*[¶]*, DECIPHER*[||]*, and OMIM*[**]*. The HPO is developed within the context of the Monarch Initiative*[††]*'*.

From this information, we have been able to generate a dataset on which we shall perform our experiments. Because the amount of information of HPO is huge, in this first approach, we are just going to use an extract of all the information available. In combination with Online Mendelian Inheritance in Man (OMIM) to distinguish amidst different types of diseases present in HPO databases, we have generated a table matching haematologic diseases and phenotypes,because the resulting information is substantial. Table IV shows an extract of the whole generated table. In the case we wish to obtain a detailed explanation of every disease shown forward, we commend the reader to visit OMIM web page and feed its search engine with the identifiers listed in Disease ID column. As an example, Disease ID 275630 corresponds to Chanarin-Dorfman syndrome.

Now, once we have presented our dataset, the next stage goes in using one of the previously mentioned techniques to retrieve all the implications that hold on it. Summing up, we are going to work over a dataset with 446 diseases, 100 different phenotypes and the set of implications that holds on it. Unfortunately, we cannot show all these implications here because of the obvious space limitations because the implications set goes beyond 6.000 implications.

In this point, we need to make an aside because from a purist view, if we calculate the so-called Duquenne-Guigues base [42] of implications that holds in context, there are 8.811 implications; so why have we pruned the set? Actually, the main feature of Duquenne-Guigues base of implications is that this base has a minimal possible number of implications among all possible bases of implications that hold in context. However, there will be several implications where there are no items that support (as stated in association rules theory) them, and usually such implications mean that set of items, contained in premise, does not occur together in context. Also, such implications include all attributes from context. Hence, the sense of this kind of implications is just theoretical and has nothing remarkable when dealing with real-world applications; that is why we ended up dismissing them. Nonetheless, the attributes present in these zero-supported implications are of course considered along the process and are fully accessible for the user to be selected.

That being said, the way on how we are going to proceed goes as follows. We are going to perform a test consisting of 1.000 simulations following to the letter the process depicted in Section 4. However, these runs will be carried out as random simulated dialogues. That is to say, we are going to conduct every dialogue by selecting phenotypes randomly from the set of available ones in each step. This strategy could give different readings. At a glance, it may deem advisable to go ahead with the random selection of phenotypes, so the reliability of the experiment is granted, and there is no possibility of inducing favourable situations for us to obtain better results. Yet the random strategy could also overshadow the benefits of our approach. On the one hand, imagine a situation where the random strategy falls into selecting phenotypes without any relation to each other. Then, the dialogue will end up quickly as there will be few diseases (or even no one) matching this phenotype's selection, and the process would finish with no possibility of applying any pruning at all. On the other hand, suppose a real dialogue where the phenotypes related to the patient share some sort of relation to each other. Then, during the dialogue, our process will be able to perform better pruning as the input phenotypes do have certain linkages; so, such a *fair-play* interaction would definitely highlight our approach. Needless to say that every result shown along with the experiments is the fruit of a statistical study [43] behind the results given from every run, so we kept the most reliable ones. Finally, in line with everything earlier, we are ready to launch the experiment.

At the end of the experiment, Figure 2 comes to clearly illustrate the results obtained for the number of steps metric after simulating 1.000 different conversations.

[§] http://www.human-phenotype-ontology.org.

[¶] http://www.orpha.net/consor/cgi-bin/index.php.

[||] https://decipher.sanger.ac.uk.

[**] http://www.omim.org.

[††] https://monarchinitiative.org.

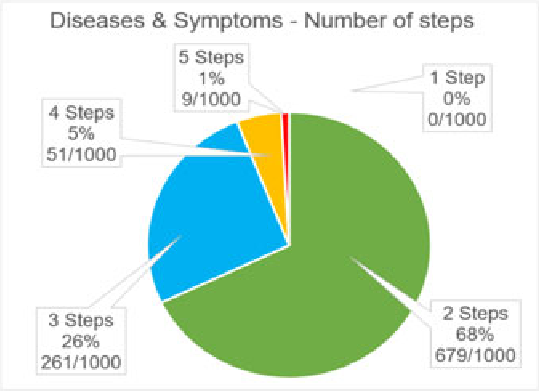

Figure 2. Number of steps metric results for the complete experiment.

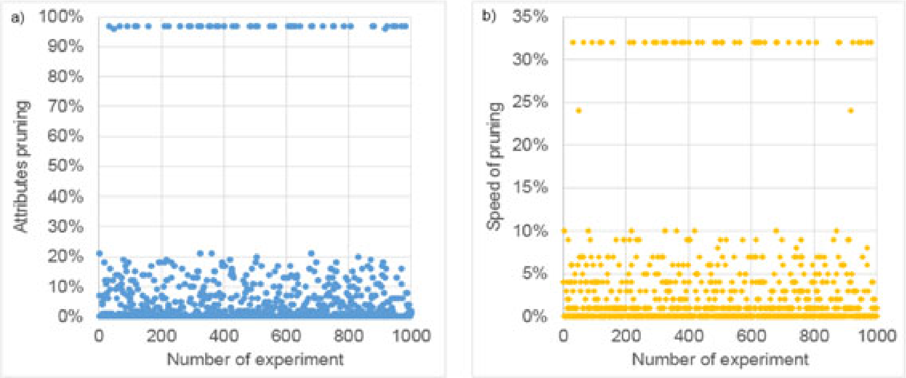

Figure 3. Attributes pruning (a) and speed of pruning (b) values for the complete experiment.

A glance is enough to easily realise that the conversational system is capable of guiding the dialogue to a suitable final diagnosis in two to three steps most of case. Keeping in mind the size of this dataset, these are encouraging results so far. There are other cases where the conversation has taken longer reaching four steps or even five, yet these cases are few and far between. Finally, only two experiments within the 1.000 runs have finished in just one step. These cases appear because the patient reveals a phenotype that appears in less than five of the diseases of the dataset and then the dialogue finishes at a glance (remember that for the random experiments, we established a limit of five or less diseases as the maximum to consider a diagnosis as a good option). Overall, we can consider that this number of step values fairly surpass previous studies [44] where the amount of input data provided is less than the one tested here.

Regarding the attributes pruning values, Figure 3(a) shows that in the vast majority of cases, the conversational system has freed us to worry about around 5–20% of attributes along the dialogue. In addition, there have been cases where the pruning did its utmost, reaching 97% of phenotypes, because the successive selected phenotypes conform a combination that brought the closure implementation to highly reduce the sets of both attributes and implications.

Speed of pruning values go hand-in-hand with the attributes pruning as it can be easily realised by matching Figure 3(a) and (b). Therefore, we can appreciate a general trend of experiments reducing attributes at an average speed of 5–10% per step and other ones achieving higher rates hovering 30–35% of attributes.

6. Conclusions and future works

We have presented here a novel application of our SL_{FD} closure algorithm in order to face the problem of the overwhelming dimensionality within the datasets. Our solution proposes a conversational process of user-driven feature selection. This work merges features of knowledge-based systems in combination with an appropriate management of implications through SL_{FD} closure. All these characteristics move us to favourably improve the diagnosis process with a high reduction in the length of the conversation and yet preserving the system's accuracy. That is not to forget the benefits we provide in terms of execution times reducing the conversational process from exponential to linear complexity.

Additionally, in the light of the results obtained over a dataset containing real information, we are of the opinion that this course of action heads the research to the right direction. The pruning rates reflect the good deeds of our approach improving the user–system interaction. Also, the number of step values encourage us to go straight on larger datasets because the actual numbers are admissible. Not forgetting, of course, the fact that results ensure 100% of accuracy.

Finally, our system constitutes a framework that can be integrated on diverse datasets and results stay admissible. This is certainly a major contribution of this work, because the possibilities it offers can reach many applications in different areas.

As future works, our results motivate a number of important directions for further research. Trying to discover which characteristics concerning the dataset (dimensionality, sparsity, etc.) are relevant to explain how the information extracted from the dataset behaves is an extended avenue for future researching tasks. In the same direction, identifying elements within the dataset, which play a more significant role amidst the others (either for being more frequent, unique, close-related to others, etc.), could help us to guide the conversation into the proper direction in a more comfortable way.

Acknowledgements

The authors thankfully acknowledge both the information provided by the Human Phenotype Ontology Consortium (HPO) and the Online Mendelian Inheritance in Man (OMIM) from the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine.

This work was supported by project TIN2014-59471-P of the Science and Innovation Ministry of Spain, co-funded by the European Regional Development Fund (ERDF).

References

1. Nahar J, Imam T, Tickle KS, Chen YPP. Computational intelligence for heart disease diagnosis: a medical knowledge driven approach. *Expert Systems with Applications* 2013; **40**(1):96–104.
2. Krajca P, Outrata J, Vychodil V. Advances in algorithms based on CBO. *Proceedings of the 7th International Conference on Concept Lattices and their applications*, Sevilla, Spain, 2010, 325–337.
3. Guerrero SE, Salamo M. Increasing retrieval quality in conversational recommenders. *IEEE Transactions on Knowledge and Data Engineering* 2012; **24**(10):1876–1888.
4. Reilly J, McCarthy K, McGinty L, Smyth B. Incremental critiquing. *Knowledge-Based Systems* 2005; **18**(4–5):143–151.
5. Trabelsi W, Wilson N, Bridge D, Ricci F. Preference dominance reasoning for conversational recommender systems: a comparison between a comparative preferences and a sum of weights approach. *International Journal on Artificial Intelligence Tools* 2011; **20**(04):591–616.
6. Jannach D, Zanker M, Fuchs M. Constraint-based recommendation in tourism: a multiperspective case study. *Information Technology & Tourism* 2009; **11**(2):139–155.
7. Chen L, Pu P. Hybrid critiquing-based recommender systems. *Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07, ACM, New York, NY, USA, 2007, 22–31.
8. Rodríguez-Jiménez JM, Cordero P, Enciso M, Mora A. Data mining algorithms to compute mixed concepts with negative attributes: an application to breast cancer data analysis. *Mathematical Methods in the Applied Sciences* 2016:4829–4845, DOI 10.1002/mma.3814.
9. Rodriguez A, Carazo JM, Trelles O. Mining association rules from biological databases. *Journal of the American Society for Information Science and Technology* 2005:493–504, DOI 10.1002/asi.20138.
10. Smyth B, McCarthy K, Reilly J, O'Sullivan D, McGinty L, Wilson DC. Case studies in association rule mining for recommender systems. *Ic-Ai*, Las Vegas, Nevada, USA, 2005, 809–815.
11. Esfandiari N, Babavalian MR, Moghadam AME, Tabar VK. Knowledge discovery in medicine: current issue and future trend. *Expert Systems with Applications* 2014; **41**(9):4434 –4463.
12. Pandey B, Mishra RB. Knowledge and intelligent computing system in medicine. *Computers in Biology and Medicine* 2009; **39**(3):215 –230.
13. Nahar J, Imam T, Tickle KS, Phoebe Chen Y-P. Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems With Applications* 2013; **40**:1086–1093.
14. Anooj PK. Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University – Computer and Information Sciences* 2012; **24**(1):27–40. arXiv:1011.1669v3.
15. Huang MJ, Chen MY, Lee SC. Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. *Expert Systems with Applications* 2007; **32**(3):856–867.
16. Imberman SP, Domanski B, Thompson HW. Using dependency/association rules to find indications for computed tomography in a head trauma dataset. *Artificial Intelligence in Medicine* 2002; **26**(1):55–68.
17. Nahar J, Tickle KS, Ali AB, Chen YPP. Significant cancer prevention factor extraction: an association rule discovery approach. *Journal of Medical Systems* 2011; **35**(3):353–367.
18. Mansingh G, Osei-Bryson KM, Reichgelt H. Using ontologies to facilitate post-processing of association rules by domain experts. *Information Sciences* 2011; **181**(3):419–434.
19. Lee DG, Ryu KS, Bashir M, Bae JW, Ryu KH. Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. *Journal of medical systems* 2013; **37**(2):98–96.
20. Yeh JY, Wu TH, Tsao CW. Using data mining techniques to predict hospitalization of hemodialysis patients. *Decision Support Systems* 2010; **50**: 439–448.
21. Hu X, Sapo M, Nenov V, Barry T, Kim S, Do DH, Boyle N, Martin N. Predictive combinations of monitor alarms preceding in-hospital code blue events. *Journal of Biomedical Informatics* 2012; **45**(5):913–921.
22. Fang R, Pouyanfar S, Yang Y, Chen SC, Iyengar SS. Computational health informatics in the big data age. *ACM Computing Surveys* 2016; **49**(1):1–36.
23. Inbarani HH, Azar AT, Jothi G. Supervised hybrid feature selection based on {PSO} and rough sets for medical diagnosis. *Computer Methods and Programs in Biomedicine* 2014; **113**(1):175 –185.
24. Vieira SM, Mendonça LF, Farinha GJ, Sousa JMC. Modified binary {PSO} for feature selection using {SVM} applied to mortality prediction of septic patients. *Applied Soft Computing* 2013; **13**(8):3494 –3504.
25. Fialho AS, Cismondi F, Vieira SM, Reti SR, Sousa JMC, Finkelstein SN. Data mining using clinical physiology at discharge to predict {ICU} readmissions. *Expert Systems with Applications* 2012; **39**(18):13158–13165.

26. Shilaskar S, Ghatol A. Feature selection for medical diagnosis: evaluation for cardiovascular diseases. *Expert Systems with Applications* 2013; **40**(10):4146–4153.
27. Mathias JS, Agrawal A, Feinglass J, Cooper AJ, Baker DW, Choudhary A. Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. *Journal of the American Medical Informatics Association* 2013; **20**(e1):e118–24.
28. Li H, Wu X, Li Z, Ding W. Online group feature selection from feature streams. *AAAI*, Bellevue, Washington, USA, 2013, 1627–1628.
29. Yu K, Wu X, Ding W, Pei J. Towards scalable and accurate online feature selection for big data. *2014 IEEE International Conference on Data Mining*, Shenzhen, China, 2014, 660–669.
30. Chaves R, Ramírez J, Górriz JM, Puntonet CG. Association rule-based feature selection method for Alzheimer's disease diagnosis. *Expert Systems with Applications* 2012; **39**(14):11766–11774.
31. Ganter B, Wille R. *Formal Concept Analysis: Mathematical Foundations* 1st ed. Springer-Verlag New York, Inc.: Secaucus, NJ, USA, 1997.
32. Cordero P, Enciso M, Mora A, de Guzmán IP. SLFD logic: elimination of data redundancy in knowledge representation. *IBERAMIA 2002: Proceedings of the 8th Ibero-American Conference on AI*, Springer-Verlag, London, UK, 2002, 141–150.
33. Armstrong WW. Dependency structures of data base relationships. *IFIP Congress*, Amsterdam, Holland, 2002, 580–583.
34. Cordero P, Enciso M, Mora A, de Guzmán IP. A tableaux-like method to infer all minimal keys. *Logic Journal of the IGPL* 2014; **22**(6):1019–1044.
35. Cordero P, Enciso M, Mora A, Ojeda-Aciego M, Rossi C. Knowledge discovery in social networks by using a logic-based treatment of implications. *Knowledge-Based System* 2015; **87**:16–25.
36. Maier D. *The Theory of Relational Databases*. Computer Science Press: Rockville, 1983.
37. Paredaens J, Bra PD, Gyssens M, Gucht DV (eds). *The structure of the relational database model*, EATCS Monographs on Theoretical Computer Science: Springer-Verlag Berlin, 1989.
38. Mora A, Cordero P, Enciso M, Fortes I, Aguilera G. Closure via functional dependence simplification. *International Journal of Computer Mathematics* 2012; **89**(4):510–526.
39. McSherry D. Minimizing dialog length in interactive case-based reasoning. *Proceedings of the 17th International Joint Conference on Artificial Intelligence, IJCAI*, Seattle, Washington, USA, 2001, 993–998.
40. Ricci F, Rokach L, Shapira B, Kantor PB (eds). *Recommender Systems Handbook*. Springer: USA, 2011.
41. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research* 2014; **42**(Database Issue):D966–D974.
42. Guigues JL, Duquenne V. Familles minimales d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences Humaines* 1986; **95**:5–18.
43. Goh TN. An information management paradigm for statistical experiments. *Quality and Reliability Engineering International* 2010; **26**(5):487–494.
44. Li H, Wu X, Li Z, Ding W. Online group feature selection from feature streams. *AAAI*, AAAI Press, Bellevue, Washington, USA, 2013.