Decision Support System for target prostate biopsy outcome prediction: clustering and FP-growth algorithm for fuzzy rules extraction

Samanta Rosati[1], Noemi Giordano[1], Enrico Checchucci[2], Sabrina De Cillis[3], Francesco Porpiglia[3] and Gabriella Balestra[1]

[1]Biolab, Department of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy

[2]Department of Surgery, Candiolo Cancer Institute, FPO-IRCCS, Candiolo, Turin, Italy

[3]Department of Oncology, Division of Urology, University of Turin, San Luigi Gonzaga Hospital, Orbassano, Turin, Italy

Abstract

An automated and data-driven rules extraction is crucial for the construction of Fuzzy Inference Systems (FIS). This work presents a method for extracting fuzzy rules based on clustering and association mining through the FP-growth algorithm. First, Self Organizing Maps are used to identify subsets of elements with similar characteristics, separately for each class. Then, the FP-Growth algorithm is applied to each cluster. Elements matching each rule are subdivided in the corresponding classes and only rules showing a predominance of elements belonging to one class are used as fuzzy rules.

The method was applied to the construction of a Decision Support System based on FIS for the target prostate biopsy outcome prediction based on six pre-bioptic variables. A dataset containing 1447 patients (824 with positive outcome, 623 with negative outcome) was used. Four and six clusters were identified for the positive and the negative class, respectively. A total of 151 rules were extracted with FP-Growth algorithm and 29 were included in the FIS. The system was able to classify 927 patients out of 1447. On the classi-fied subjects, it reached a sensitivity of 87.5% and a specificity of 58.8%.

Keywords

Rule extraction, Fuzzy Logic, Association Mining, FP-growth, Decision Support System

1. Introduction

Ever since their first implementation, Fuzzy Inference Systems (FIS) have proven a valuable tool to perform classification tasks in complex environments. Their main advantage resides in their capability of dealing with uncertainty, which is typical of biological systems and is often difficult to model in real life applications [1, 2]. In the latest decade, indeed, FIS were often used in the medical field for the implementation of decision support systems, with promising results [3, 4].

The main critical aspect in the design of a FIS is the construction of the fuzzy rules. On their

AIxIA 2021 SMARTERCARE Workshop, November 29, 2021, Milan, IT

samanta.rosati@polito.it (S. Rosati); noemi.giordano@polito.it (N. Giordano); checcu.e@hotmail.it (E. Checchucci); sabrinatitti.decillis@gmail.com (S. D. Cillis); francesco.porpiglia@unito.it (F. Porpiglia); gabriella.balestra@polito.it (G. Balestra)

0000-0003-0620-594X (S. Rosati); 0000-0002-9265-6538 (N. Giordano); 0000-0001-7901-5608 (E. Checchucci); 0000-0002-7798-6926 (S. D. Cillis); 0000-0002-0752-4857 (F. Porpiglia); 0000-0003-2717-648X (G. Balestra)

basis, the combination of the activations of specific membership functions in the input variables leads to a membership degree for the output variable MFs. In several applications, fuzzy rules are typically identified according to expert knowledge. Nevertheless, it was often proved that FIS based on expert knowledge lack of generalizability in the case of complex systems and are uncapable of accurately matching the wide range of combinations that a high number of features may convey [1, 2].
In this context, research on data-driven approaches for the definition of fuzzy rules is currently of high interest. In this work, we present a method for extracting fuzzy rules from a dataset based on clustering and association mining through the FP-growth algorithm. Our method allows to find underlying patterns within patients with a similar outcome, cutting out the need to integrate any expert knowledge.
We applied our method to the design of a FIS-based Decision Support System (DSS) for the prediction of the outcome of target Prostate Biopsy (PB) based on six pre-bioptic variables. The final goal of the implemented system is the pre-selection of ideal candidates for target PB, thus reducing the high number of unneeded biopsies that are performed nowadays.

2. Materials and Methods

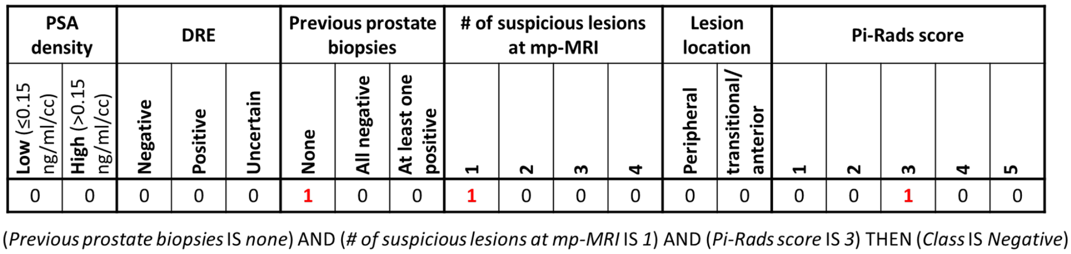
2.1. Fuzzy Rule Extraction

The method for the extraction of fuzzy rules is made of four steps: (1) clustering using Self Organizing Maps (SOMs) and hierarchical clustering; (2) dataset binarization; (3) rule extraction using the FP-growth algorithm; and (4) check of class predominance and rules input into the FIS.

Step 1: Clustering. The first step consists of dividing the dataset into groups according to the class and in applying clustering to each group in order to identify subgroups with similar characteristics. The SOMs combined with agglomerative hierarchical clustering are employed to this scope. In particular, a SOM is trained for each group of elements. The network parameters (dimension, neighborhood size, training steps and neuron distance function) are tuned according to the problem and the dataset characteristics. Once the SOM is trained, the agglomerative hierarchical clustering with the complete linkage method was applied to the neurons weights to obtain clusters of neurons. Finally, the elements belonging to neurons of the same cluster are assembled to obtain subgroups of elements with similar characteristics (clusters).

Step 2: Dataset Binarization. Since the FP-Growth algorithm that is applied to the next step works only on binary data, each non-binary original variable (continuous, categorical or discrete) must be transformed into a binary one. To this scope, each continuous or discrete variable is divided into intervals and each couple variable-interval is associated to a new binary variable: for each original value that the variable assumes for a given element, a “1” is set in the corresponding couple variable-interval; all the other couples related to the same variable are set to “0”. Intervals must be defined according to the fuzzy input variables used for the FIS: each membership function of a fuzzy variable must correspond to an interval in the binary matrix. This procedure is applied to every cluster to obtain a binary matrix for each of them.

Step 3: Rule Extraction. To automatically extract the rules, the FP(Frequent Patterns)-Growth algorithm is used, that allows frequent pattern generation based on a compressed representation of the items in a dataset called FP-tree [5]. The FP-Growth is applied to the binary matrix associated to each cluster (the class variable is not used in this phase) and returned the list of the frequent itemsets mined from it. The algorithm parameters (support, confidence and lift) are tuned according to the specific problem. Finally, each itemset is transformed into a rule: each item in the itemset is transformed into a couple variable-interval and used as antecedent of the rule; all antecedents are linked with the AND operator; the rule consequent is assigned to the class which the elements in the considered cluster belongs to. An example of rule generation from an itemset containing 3 items and extracted from a cluster belonging to the negative class is reported in Fig. 2.1. Itemsets containing only one item are not used for the rule extraction since they would produce too generic rules (containing only one antecedent). Once the rule extraction is performed for every cluster, duplicated rules (i.e. having equal antecedents and consequent) are removed from the list.



PSA density		DRE			Previous prostate biopsies			# of suspicious lesions at mp-MRI				Lesion location		Pi-Rads score				
Low (≤0.15 ng/ml/cc)	High (>0.15 ng/ml/cc)	Negative	Positive	Uncertain	None	All negative	At least one positive	1	2	3	4	Peripheral	transitional/ anterior	1	2	3	4	5
0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0

IF (*Previous prostate biopsies* IS *none*) AND (*# of suspicious lesions at mp-MRI* IS *1*) AND (*Pi-Rads score* IS *3*) THEN (*Class* IS *Negative*)

Figure 1: Example of rule generation from a cluster of elements belonging to the negative class.

Step 4: Check of class predominance and rules input into the FIS For each rule obtained from the previous step, the class predominance is assessed. First, the number of elements matching the rule is counted for each class separately. Then, the rule predominance is calculated as the ratio between the number of elements belonging to one class with respect to the number of elements belonging to the other classes. Only rules showing a predominance of elements belonging to one class are kept and input into the FIS. In this way, rules that are not class-specific are removed. A threshold on the predominance is set according to the specialization that the system aims to reach: higher thresholds lead to rules that are more specific for a given class; however, this could lead to a lower number of elements matching the rules and thus classified by the FIS.

2.2. Validation

The described method was validated through its usage for the construction of a FIS-based DSS for the prediction of the outcome of target PB. In particular, the method was applied on a sample male population counting 1447 patients, whose data were recorded from March 2014 to

December 2019 and retrospectively analyzed. For each patient, six pre-bioptic features were taken into account, namely PSA density, digital rectal examination (DRE), previous PBs, number of suspicious lesions at mp-MRI, lesion location and Pi-Rads score. 824 patients had a positive biopsy outcome and were classified as class "1", whereas 623 patients had a negative biopsy outcome and were classified as class "0".

The FIS was made of 6 input variables, corresponding to the pre-bioptic features and modelled with a set of trapezoidal or triangular Membership Functions (MFs), according to the variable, and 1 output variable associated to the patient class and modelled with two triangular MFs (negative and positive).

The method results for each step were reported and the final performances were indirectly assessed by means of the performance of the FIS in classifying the entire sample population. All steps were implemented using MATLAB® release R2021a (The MathWorks Inc., Natick, MA, USA).

3. Results and Discussion

The clustering of the two groups of patients was performed using two SOMs with dimension 4x4 and 5x5 for the positive and negative group, respectively. Four clusters of positive patients and 6 clusters of negative patients were identified using the agglomerative hierarchical clustering, highlighting a higher heterogeneity in the latter group, even though it is smaller than the positive group.

The dataset binarization allows to obtain 19 binary couples variable-interval: 2 intervals were identified for the PSA density (≤0.15 ng/ml/cc; >0.15 ng/ml/cc), 3 intervals for the DRE (negative, positive, uncertain) and the previous PBs (none, all negative, at least one positive), 4 intervals for the number of suspicious lesions at mp-MRI (1, 2, 3, 4 suspicious lesions), 2 intervals for the lesion location (peripheral, transitional/anterior) and 5 intervals for the Pi-Rads score (1, 2, 3, 4, 5).

The FP-Growth algorithm was applied to binary matrix of each cluster setting the following parameters: support = 0.25; confidence = 0.5; lift = 1. These values were selected after a tuning phase in which the performance of the final FIS was evaluated for each set of extracted rules. A total of 151 rules were extracted with FP-Growth algorithm after the removal of itemsets containing only one item and duplicated rules: 99 rules for the negative class and 52 rules for the positive class. Again, this result confirms the greater heterogeneity of negative patients, since more rules are needed to describe this population. This intermediate result is neither obvious nor evident from the analysis of the entire dataset: a higher variability is usually expected in the positive population, due to differences in the tumor characteristics (e.g. aggressiveness).

We set the threshold for the class predominance equal to 3: this means that only rules matching a number of positive patients at least triple with respect to the negative ones (or vice-versa) were included in the FIS. This led to a selection of 29 rules: 21 rules for the classification in the negative class and the remaining 8 rules for the positive class. The list of the 29 rules is reported in Table 1.

Using only 29 rules, the FIS was able to classify 927 patients out of 1447 (405 negative and 522 positive). On the classified subjects, it reached a total accuracy of 75%, with sensitivity =

Table 1
Rules inserted into the FIS.

N	Antecedent	Consequent
1	IF (lesion location IS periph) AND (Pi-Rads IS 3)	THEN (class IS n)
2	IF (N of lesions IS 1) AND (Pi-Rads IS 3)	THEN (class IS n)
3	IF (lesion location IS periph) AND (N of lesions IS 1) AND (Pi-Rads IS 3)	THEN (class IS n)
4	IF (previous PBs IS all negative) AND (Pi-Rads IS 3)	THEN (class IS n)
5	IF (previous PBs IS none) AND (N of lesions IS 1) AND (Pi-Rads IS 3)	THEN (class IS n)
6	IF (DRE IS negative) AND (Pi-Rads IS 3)	THEN (class IS n)
7	IF (DRE IS negative) AND (lesion location IS periph) AND (Pi-Rads IS 3)	THEN (class IS n)
8	IF (DRE IS negative) AND (N of lesions IS 1) AND (Pi-Rads IS 3)	THEN (class IS n)
9	IF (DRE IS negative) AND (N of lesions IS 1) AND (lesion location IS periph) AND (Pi-Rads IS 3)	THEN (class IS n)
10	IF (DRE IS negative) AND (previous PBs IS none) AND (Pi-Rads IS 3)	THEN (class IS n)
11	IF (DRE IS negative) AND (previous PBs IS none) AND (N of lesions IS 1) AND (Pi-Rads IS 3)	THEN (class IS n)
12	IF (DRE IS negative) AND (previous PBs IS none) AND (N of lesions IS 1) AND (lesion location IS periph) AND (Pi-Rads IS 3)	THEN (class IS n)
13	IF (PSA density IS low) AND (Pi-Rads IS 3)	THEN (class IS n)
14	IF (PSA density IS low) AND (lesion location IS periph) AND (Pi-Rads IS 3)	THEN (class IS n)
15	IF (PSA density IS low) AND (N of lesions IS 1) AND (Pi-Rads IS 3)	THEN (class IS n)
16	IF (PSA density IS low) AND (N of lesions IS 1) AND (lesion location IS periph) AND (Pi-Rads IS 3)	THEN (class IS n)
17	IF (PSA density IS low) AND (previous PBs IS none) AND (Pi-Rads IS 3)	THEN (class IS n)
18	IF (PSA density IS low) AND (DRE IS negative) AND (Pi-Rads IS 3)	THEN (class IS n)
19	IF (PSA density IS low) AND (DRE IS negative) AND (lesion location IS periph) AND (Pi-Rads IS 3)	THEN (class IS n)
20	IF (PSA density IS low) AND (DRE IS negative) AND (N of lesions IS 1) AND (Pi-Rads IS 3)	THEN (class IS n)
21	IF (PSA density IS low) AND (DRE IS negative) AND (previous PBs IS none) AND (Pi-Rads IS 3)	THEN (class IS n)
22	IF (lesion location IS periph) AND (Pi-Rads IS 5)	THEN (class IS p)
23	IF (previous PBs IS none) AND (Pi-Rads IS 5)	THEN (class IS p)
24	IF (previous PBs IS none) AND IF (lesion location IS periph) AND (Pi-Rads IS 5)	THEN (class IS p)
25	IF (DRE IS positive) AND (lesion location IS periph)	THEN (class IS p)
26	IF (DRE IS positive) AND (N of lesions IS 1)	THEN (class IS p)
27	IF (DRE IS positive) AND (N of lesions IS 1) AND (lesion location IS periph)	THEN (class IS p)
28	IF (PSA density IS high) AND (previous PBs IS none)	THEN (class IS p)
29	IF (PSA density IS high) AND (previous PBs IS none) AND (lesion location IS periph)	THEN (class IS p)

87.5% and specificity = 58.8%. This means that, although the limited number of rules extracted for the positive class (only 8 rules), they are able to correctly describe the characteristics of more than the 87% of patients that will result positive to the biopsy. On the other hand, the lower performance on the negative class (specificity =58.8%) could be explained with the high heterogeneity that emerged for these patients in the previous steps. This problem could be faced, for example, by clustering these patients using a higher number of clusters, in order to further reduce the intra-cluster variability. Another criticality emerging from these results could be the presence of patients that are not classified by the system. However, from the clinical point of view, it is possible to consider these subjects (218 negative and 302 positive) as "at risk" and perform the biopsy.

4. Conclusions

In this study we present a methodology that combines clustering and association mining for the extraction of a set of fuzzy rules in an automated, data-driven way. We presented the results of the methodology application for the extraction of rules to be included in a FIS-based DSS for the prediction of the outcome of target prostate biopsy. Our results, although preliminary, proved the effectiveness of the methodology, since a sensitivity of 87.5% was reached by our system. From the intermediate results of the method it is possible to mine new knowledge about the dataset heterogeneity.

References

[1] O. Gorgulu, A. Akilli, Use of fuzzy logic based decision support systems in medicine, Studies on Ethno-Medicine 10 (2016) 393–403.

[2] E. Hüllermeier, Fuzzy methods in machine learning and data mining: Status and prospects, Fuzzy sets and Systems 156 (2005) 387–406.

[3] Fuzzy logic applied to a Patient Classification System, IEEE, 2013.

[4] Basographic gait impairment score: A fuzzy classifier based on foot-floor contact parameters, IEEE, 2014.

[5] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, ACM sigmod record 29 (2000) 1–12.