Big Data Based Extraction of Fuzzy Partition Rules for Heart Arrhythmia Detection: a Semi-Automated Approach

Omar Behadada[1], Marcello Trovati[2], Chikh MA[1], Nik Bessis[2]
[1] Biomedical Engineering Laboratory, Department of Biomedical Engineering, Faculty of technology, University of Tlemcen, Algeria
[2]School of Computing and Mathematics, University of Derby, UK
o_behadada@mail.univ-tlemcen.dz; M.Trovati@derby.ac.uk; mea_chikh@mail.univ-tlemcen.dz; N.Bessis@derby.ac.uk

***Abstract* — In this paper, we introduce a novel method to define semi-automatically fuzzy partition rules to provide a powerful and accurate insight into cardiac arrhythmia. In particular, we define a text mining approach applied to a large data-set consisting of the freely available scientific papers provided by PubMed. The information extracted is then integrated with expert knowledge, as well as experimental data, to provide a robust, scalable, and accurate system, which can successfully address the challenges posed by the management and assessment of big data in the medical sector. The evaluation we carried out shows an accuracy rate of 93% and interpretability of 0.646, which clearly shows that our method provides an excellent balance between accuracy and system transparency. Furthermore, this contributes substantially to the knowledge discovery and offers a powerful tool to facilitate the decision making process.**

Keywords: Knowledge Discovery, Text Mining, Fuzzy Logic, Cardiac Arrhythmia, Big Data, Data Analytics

I. INTRODUCTION

Cardiovascular diseases are one of the most worrying health issues, and the largest cause of mortality in the world, based on the World Health Report 2013 [1]. Thus, low-cost and high-quality cardiac assessment offers a very valuable challenge. Furthermore, the availability of a huge amount of information created by the continuously development of big data methods and techniques, provides new challenges as well as new opportunities in this field.

The detection of cardiac arrhythmia is a very interested area, since *premature ventricular contraction* (PVC) is an effective predictor of sudden death. Several studies, over the past decade have focused on methods and algorithms for detection and significance of cardiac arrhythmias, aiming to achieve a good classification rate.

More specifically, many classification methods have been used in the field such as Bayesian classifiers, decision trees, neural, and rule based learners [2]. However, the classification methods with good classification rates usually have a low degree of interpretability preventing the user (such as a cardiologist) from fully taking advantage of such method.

The *interpretability* of any knowledge based system is crucial, especially when dealing with big data applications [3]. In fact, this ensures an effective decision making progress by producing interpretable knowledge, which can easily be maintained and assessed. The acquisition of expert knowledge is a complex, yet essential task due to its inherent accuracy, if carried out by human intervention. However this tends to be inefficient when dealing with big data-sets and furthermore, such knowledge contains an unconscious component, which is hard to formalise [4]. Alternatively, knowledge can also be defined by analysing information extracted from experimental big data, which is likely to provide an accurate insight into the different parameters [5]. In particular, there is a wealth of algorithms and machine-learning techniques for model identification, which are based on the properties of accuracy indices, which can be applied to the knowledge induction process [6].

Another valuable source of knowledge is based on articles and texts published in scientific journals, which include the most critical knowledge, and in which many experts share their results, analysis, and experiences. However, since such textual information is typically very large, if not huge, scientists are faced with great amount of information, which poses a huge computational and implementation challenge. In modern medicine, large amounts of data are generated, but there is a widening gap between data acquisition and data comprehension. It is often impossible to process all of the data available and to make a rational decision on basic trends. Thus, there is a growing need for intelligent data analysis, such as data mining, to facilitate the creation of knowledge to support clinicians in decision making. In fact, data mining approaches can be used in such databases, to improve classification tasks.

In this paper, we introduce a novel method to semi-automatically identify fuzzy partition rules applied to cardiac arrhythmia detection, which combines an automated

information extraction from textual sources with expert elicitation to create a robust, scalable, and accurate knowledge based system, which provides a crucial insight into arrhythmia detections from big data information sources.

The paper is structured as follows: in the rest of this section we discuss the relevant medical background, in Section III we introduce the text mining method to extract information for the generation of fuzzy partitions, which are subsequently analysed and evaluated in Sections IV, V, VI and VII. Finally in Section VIII, we discuss the main findings and future research directions.

A. *Medical Context*

Electrocardiogram (ECG) reflects the activity of the central blood circulatory system, which can provide extensive information on the normal and pathological physiology of heart activity. See Figure 1 for an example of the main features of ECGs. As a consequence, it is an important non-invasive clinical tool for the diagnosis of heart diseases [7]. Early and quick detection and classification of ECG arrhythmia is important, especially for the treatment of patients in the intensive care unit [8]. For over four decades, computer-aided diagnostic (CAD) systems have been used in the classification of the ECG resulting in a huge variety of techniques. Included in these techniques are multivariate statistics, decision trees, fuzzy logic, expert systems and hybrid approaches [8]. In designing of CAD system, the most important step is the integration of suitable feature extractor and pattern classifier such that they can operate in coordination to make an effective and efficient CAD system [9].

B. *Big Data in Medical Information Retrieval*

The medical sector has always generated large amounts of data, based on patient record, regulatory requirements, etc. [10]. All the information created by the above data provides a valuable opportunity to further improve the state-of-the-art tools available to clinicians, as well as to provide scalable, robust and efficient methods to extract, assess and manage such information [11].

Furthermore, the digitalisation of medical data-sets with the implementation of big data techniques, has created further benefits. These include the detection of diseases at earlier stages resulting to more effective and successful treatments, as well as the management of specific individual and population medical information. Specific scenarios can also be predicted and estimated based on large amounts of historical data, combined with real-time information to determine and assess their crucial properties [12].

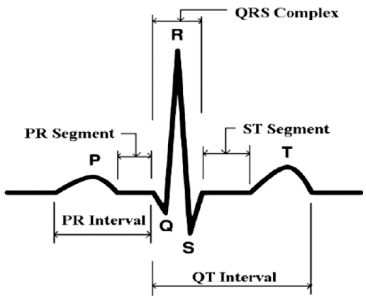

Figure 1: Standard ECG beat

II. Description of the Method

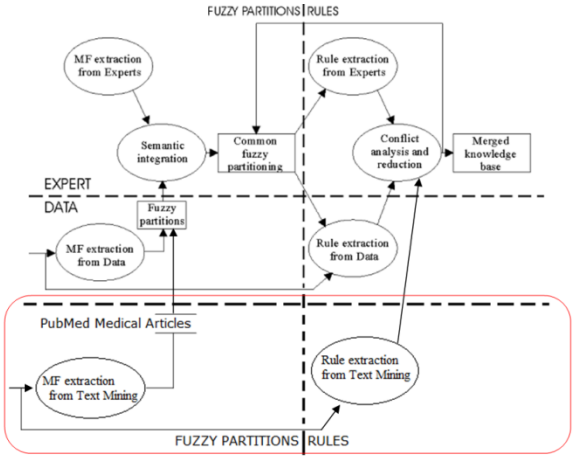

Figure 2: Overall structure of the knowledge extraction process

The method we are proposing, and in particular the overall structure of the extraction process from expert knowledge, data, and textual information is depicted in Figure 2.

More specifically, fuzzy logic as modelling platform allows us to merge and manage those three types of knowledge, where the fuzzy partition design aims to define the most influential variables, according to the above knowledge. An important part of this process is the rule base definition and integration, where the expert is invited to make a description of the system behaviour, expressing his/her system knowledge as linguistic rules (*expert rules*). Furthermore, rules are induced from data (*induced rules*) according to the common universe of fuzzy partition. Both types of rules use the same linguistic terms defined by the same fuzzy sets. As a consequence, rule comparison can be done at the linguistic level and subsequently, both types of rules are merged into a unique knowledge base. As part of the process, the expert can provide complete or partial information about the linguistic variables. Additionally, several algorithms can be used to create fuzzy partitions from data, or *induced partitions*, and linguistic constraints are superimposed to the fuzzy partition definition, in order to ensure their interpretability. The result is the definition of a common universe for each of the variables, according to both expert knowledge and data distribution.

III. Automated Extraction of Fuzzy Partition Rules From Text

Text mining (TM) [13] is a branch of computer science, which aims to accurately extract, identify and analyse information and semantic properties from text sources. Even though there has been steady and successful progress in addressing the above challenges, TM research is still very much expanding to provide further state-of-the-art tools to improve accuracy, scalability and flexibility. The extraction of information from textual sources is typically a complex task due to the ambiguous nature of human language. In fact, depending of the general context and the given semantic information, a variety of text mining techniques can be used, which in general depend on the type of data and their structure [13].

In this paper, we apply a grammar based text extraction, based on *text patterns*, which relies on a set of rules identifying sentences with a determined structure. More specifically, we consider text patterns of the form `(NP, verb, NP)`, where `NP` refers to the noun phrases, and `verb` to the linking verb [13]. For example, sentences such as "*PVCs can be related to electrolytes*" are identified to extract a relationship between *PVCs* and *electrolytes*. The effectiveness of this approach is fully exploited when syntactic properties of a sentence are investigated, by using suitable parsing technology [14]. In particular, the syntactic roles of the different phrasal components are essential in extracting the relevant information, and they can contribute towards a full understanding of the type of relationship. Furthermore, we also apply basic sentiment analysis, which aims to identify the mood described by text fragments based on specific keywords [15]. Table 1 shows a small selection of such keywords used in our approach. We mined all the articles in journals freely available from PubMed [16], a very large data base containing biomedical literature, as follows:

- We identified articles from the above journals containing the following keywords:
    - Premature ventricular contractions, or PVCs
    - Premature ventricular complexes
    - Ventricular premature beats
    - Extrasystoles
- The identified articles were first lexically and syntactically analysed via the Stanford Parser [14].
- Subsequently, a grammar-based extraction identifies the relevant information based on the above keywords as well as on sentiment analysis [14]. More specifically, only sentences with one or more of the above keywords, including those in Table 1, in the `NPs` will be extracted.

Table 1: A selection of keywords used

Negative Keywords	Positive Keywords	Uncertain Keywords
Bad	Satisfactory	Unpredictable
Negative	Enhancing	Possible
Underestimate	Advantage	Somewhat
Unsafe	Benefit	Precautions
Unwelcome	Good	Speculative
Tragic	Excellent	Confusing
Problematic	Great	Fluctuation

A. Text Mining Extraction Results

Table 2: Example of relation extraction

Keywords in Relation Extraction	Sentences identified
'PVC', 'PVCs', 'imbalances'	'PVCs can be related to electrolyte or other metabolic imbalances'
'PVC', 'death'	'70 mmol/L and T2DM significantly increases risk of PVC and sudden cardiac death, the association between sMg and PVC may be modified by diabetic status'
'premature ventricular complexes', 'PVC', 'PVCs', 'atrial', 'ventricular', 'beat', 'missed', 'premature'	'The system recognises ventricular escape beats, premature ventricular complexes (PVC), premature supraventricular complexes, pauses of 1 or 2 missed beats, ventricular bigeminy, ventricular couplets (2 PVCs), ventricular runs (> 2 PVCs), ventricular tachycardia, atrial fibrillation/flutter, ventricular fibrillation and asystole'

The output of the extraction consists of the set of keywords found in each text fragment (usually a sentence), which was also extracted, see Table 2 for an example. A full assessment of this type of information extraction from text goes beyond the scope of this paper, since it specifically addresses issues that are not directly relevant in this context. However, we considered a small evaluation consisting of two randomly chosen papers [17] [18], from those identified above. The automatic extraction was then compared with a manual one, which produced a recall of 71% and a precision of 84%. In future research, we will investigate a more specialised set of keywords, as well as a larger set of text patterns compared to those utilised in this paper, to improve the above measures.

IV. Data preparation

The patients who have been considered in the experiments, taken from MIT-BIH [19], are shown in Table 3. The R peaks of the ECG signals were detected using the Tompkins algorithm [20], which is an online real time QRS detection algorithm. This algorithm reliably detects QRS complex using slop, amplitude, and width information. This algorithm automatically adjust thresholds and parameters periodically to the standard 24h MIT-BIH arrhythmia

database, this algorithm correctly detects 99.3 percent of QRS complex.

From patients with cardiac arrhythmia, taken from MIT-BIT database, we chose only patients with three conditions, namely premature ventricular contraction beats (PVC) premature arterial contraction beats (PAC) and premature junctional contraction beats (PJC), since they provide the best quality of records, and more specifically PVC is a predictive element of the CA sudden death.

Table 3: Evaluation data taken from the MIT-BIH database.

Record	N	A	J	V
101	1860	3	-	-
103	2082	2	-	-
104	163	-	-	2
105	2526	-	-	41
106	1507	-	-	520
107	-	-	-	59
108	1739	4	-	17
109	-	-	-	38
111	-	-	-	1
112	2537	2	-	-
113	1789	-	-	-
114	1820	10	2	43
115	1953	-	-	-
116	2302	1	-	109
117	1534	1	-	-
118	-	96	-	16
119	1543	-	-	444
121	1861	1	-	1
122	2476	-	-	-
123	1515	-	-	3
124	-	2	29	47
200	1743	30	-	826
201	1625	30	1	198
202	2061	36	-	19
203	2529	-	-	444
205	2571	3	-	71
207	-	107	-	105
208	1586	-	-	992
209	2621	383	-	1

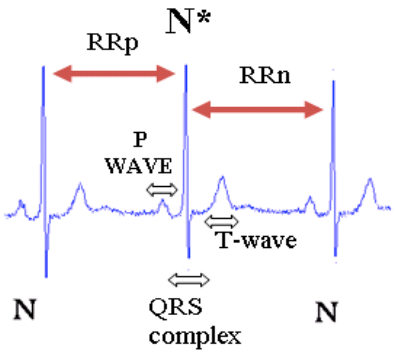

Figure 3: Standard ECG beat

Dataset:

Class	*Normal*	*PVC*	*PAC*	*PJC*
Number of samples	60190	6709	2130	83

A. Feature Selection

The feature vector, which is used for recognition of beats, has been selected as follows:

- the R-R interval of the beat RRp (calculated as the difference between the QRS peak of the present and previous beat),
- the ratio = RR1-to-RRo (RRn is calculated as the difference between the QRS peak of the present and following beat see Figure 3), and
- the QRS width (calculated according to the Tompkins algorithm [20]).

In this way, each beat is stored as 3-element vector. Table 4 provides the most relevant parameters used in this type of analysis.

Table 4: The various descriptors

Attributes	Meaning
RR precedent: RR0	The distance between the peak of this beat R and R of the peak beat precedent
RR next : RRn	RRn between the peak the present R and beat the peak of R beat following
QRS *complex*	Beginning of the Q-wave and the end of the S wave
Comp	The ratio RR0/RRs
PP	Pic to pic of the R wave of the QRS complex
Energy	Energy of the QRS complex

V. Fuzzy Partition Design

This section covers the left most parts, as depicted in Figure 2, i.e. those related to membership functions extraction from both the expert's input and experimental data. Note that the initial step considers the extraction from the former, and subsequently some approaches for membership function design from data are introduced.

When defining expert knowledge, we make the assumption that the linguistic variables of the system are sufficiently known. More specifically, experts may identify specific properties, such as a domain of interest within a physical range, who would be subsequently facilitated in the decision process by a given, and possibly small, number of linguistic terms. In this paper, as discussed above, we assume that a minimum information on membership function definition is available, which includes the definition of universes, number of terms, and, sometimes, prototypes of linguistic labels (modal points) [21]. Note that this is a reduced version of the interval estimation method [22], as the interval is reduced to a single point. Furthermore, if additional information is provided by the expert, this can be integrated into the system.

The knowledge base is split into two main parts, the *data base* (DB) and the *rule base* (RB). In particular, the former is defined by the description of the linguistic variables such as number, range, granularity, membership functions, as well as their normalisation functions. As discussed in [21], in most of the existing approaches, which focus on the generation from data of the fuzzy partitions, the automatic design of the data base is one of the most important steps in the definition of the overall knowledge base. However, the method we are proposing in this paper consists of rules that integrate expert as well as data-based knowledge.

The automatic generation of fuzzy partitions is based on the definition of the best shapes of the membership functions, in terms of the optimal number of linguistic terms in the fuzzy partitions, and the location of the fuzzy sets within the universe of the variable.

As discussed in [23], in this paper we follow an approach, which includes:

1. A non-supervised clustering process is performed to address the extraction of the DB from the available data set as part of the preliminary design,
2. An embedded basic learning method, which derives the DB.

Simultaneous design [23] is another method, which cannot be successfully applied in this context. In fact, it usually generates a far more complex process in which the computational effort involved is not fully exploited since only the fuzzy partitions are considered. In addition, our interpretability requirements require some specific properties for the partitions. In fact, techniques generating multidimensional clusters cannot be successfully applied, since only one-dimensional membership functions are required. On the other hand, it is feasible to apply a one-dimensional optimisation technique if it includes some semantic constraints.

In the case of embedded design, when search techniques are used in the design of the DB, it is essential to include appropriate interpretability measures in the objective function, to provide suitable and optimal solutions. These include measures of completeness, consistency, compactness, or similarity. Finally, at the end of the embedded design process, only fuzzy partitions are considered, which will subsequently lead to the creation of fuzzy rules.

A. Criteria for the Evaluation of Fuzzy Partitions

The evaluation of fuzzy partitions is based on both linguistic properties and partitioning properties [21]. The former influences the shape of the fuzzy sets, as well as the relations between fuzzy sets corresponding to the same variable. The latter, on the other hand, focus on the data from the partitions that have been generated, and more importantly, on their level of matching with the partitions derived from data. This is not the case for linguistic properties since their assessment does not involve data.

In the remaining section, we will give a brief overview of some features of partitioning, as introduced in [21]. In particular, similar to the above mentioned paper, we only focus on

1) Methods based on the data distribution, which excludes methods based on an input-output relation.
2) Methods typically applied in unsupervised clustering, and not in supervised clustering.
3) The assignment of data-elements to each of the item of the partition.

Using the same notation as in [21], let u_{ik} be the degree of membership of the k-th element of the data set to the i-th element of the fuzzy partition. The partition coefficient is

$$PC = \frac{\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^2}{n}$$

and the partition entropy is

$$PE = -\frac{1}{n}\left\{\sum_{k=1}^{n}\sum_{i=1}^{c}\left[u_{ik}\log(u_{ik})\right]\right\}$$

where c is the number of elements of the fuzzy partition, and n is the cardinality of the set of data. Furthermore, Chen [24] recently introduced the following index measure

$$Ch = \frac{1}{n}\sum_{k=1}^{n}\max_{i} u_{ik} - \frac{2}{c(c-1)}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}\frac{1}{n}\sum_{k=1}^{n}\min(u_{ik}, u_{jk})$$

An efficient partition should minimise the entropy and maximise the coefficient partition and the Chen index [21].

Table 5: RRo fuzzy partition quality (3 labels)

Partition	*Partition Coefficient(max)*	*Partition Entropy(min)*	*Chen Index(max)*
HFP	0.77513	0.33406	0.77094
Regular	0.69936	0.47483	0.74665
K-means	0.79948	0.30842	0.80794
Expert &TM	0.78262	0.32504	0.78647

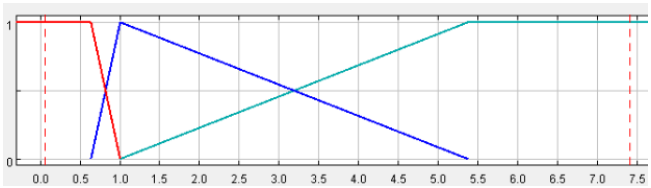

Figure 4: Fuzzy partition RR0 from K-means algorithm

Table 6: RRs fuzzy partition quality (3 labels)

Partition	*Partition Coefficient(max)*	*Partition Entropy(min)*	*Chen Index*
HFP	0.77378	0.33614	0.76980
Regular	0.69705	0.47751	0.74360
K-means	0.77121	0.34723	0.77941
Expert &TM	0.78300	0.32441	0.78668

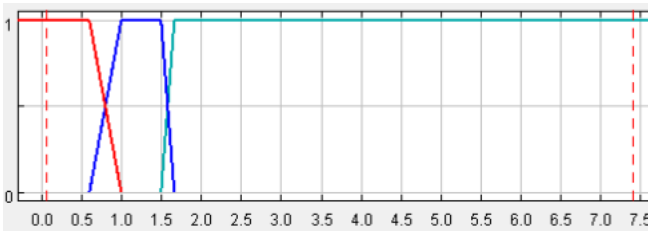

Figure 5: Fuzzy partition RRs from Expert & TM

Table 7: QRS fuzzy partition quality (3 labels)

Partition	*Partition Coefficient(max)*	*Partition Entropy(min)*	*Chen Index(max)*
HFP	0.76812	0.34495	0.76312
Regular	0.66966	0.50649	0.70024
K-means	0.82214	0.26975	0.82707
Expert &TM	0.84540	0.22046	0.82512

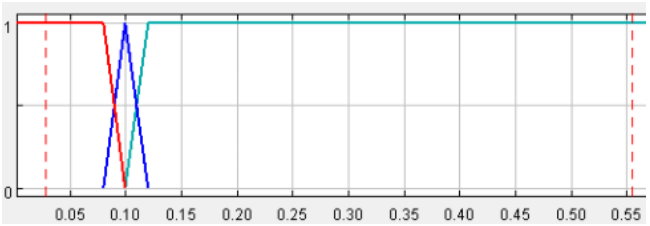

Figure 6: Fuzzy partition QRS from Expert & TM

Table 8: COMP fuzzy partition quality (3 labels)

Partition	*Partition Coefficient(max)*	*Partition Entropy(min)*	*Chen Index(max)*
HFP	0.68847	0.45294	0.67092
Regular	0.71224	0.45990	0.75945
K-means	0.89473	0.18398	0.90993
Expert &TM	0.81362	0.29566	0.82971

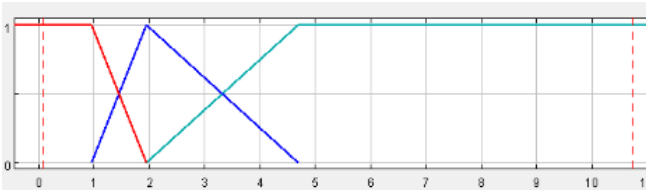

Figure 7: Fuzzy partition COMP from K-means algorithm

Table 9: PP fuzzy partition quality (3 labels)

Partition	*Partition Coefficient(max)*	*Partition Entropy(min)*	*Chen Index(max)*
HFP	0.68847	0.45294	0.67092
Regular	0.71224	0.45990	0.75945
K-means	0.89473	0.18398	0.90993
Expert &TM	0.81362	0.29566	0.82971

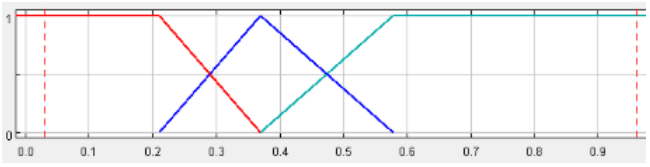

Figure 8: Fuzzy partition PP from kmeans algorithm (x10^3)

Table 10: Energy fuzzy partition quality (2 labels)

Partition	*Partition Coefficient(max)*	*Partition Entropy(min)*	*Chen Index(max)*
HFP	0.78374	0.31852	0.66240
Regular	0.52352	0.66896	0.17696
K-means	0.77454	0.33891	0.67045

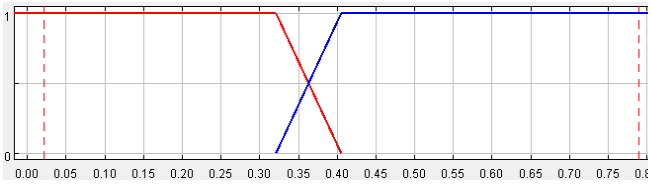

Figure 9: Fuzzy partition energy from HFP algorithm

The most important part of the knowledge-based systems is the reasoning mechanism that induces decision rules. Since the fuzzy partition is at the core of induction methods in fuzzy logic, in this paper we proposed to initialize the fuzzy partitions by two approaches: the first, a purely automatic method of induction (*K-means, HFP* and *regular*) and the approach resulting from information extraction from textual sources, as discussed in Section III, to compare between the different methods.

Subsequently, we have established linguistic terms to construct the rules and modal point.

VI. Rule Base Generation

The process of generating rules from data is called *induction*, which aims to produce general statements, expressed as fuzzy rules in our case, valid for the whole set, from partial observations. The observed output, for each sample item, is part of the training set allowing supervised training procedures. Many methods are available in the fuzzy logic literature [25], but we are only interested in those ones, which generate rules sharing the same fuzzy sets. Thus, we have chosen the methods, which are implemented in Fispro [25] and they are used by KBCT [1], as they can be run with previously defined partitioning.

A. Knowledge Base Accuracy

In order to obtain an accuracy measure, we need to compare the inferred output with the observed one in a real system. In classification systems, the most common index is defined as the number of misclassified cases. We will consider the three following indices:

- Unclassified cases (*UC*): Number of cases from data set that do not fire at least one rule with a certain degree.
- Ambiguity cases (*AC*): Number of remaining cases for which the difference between the two highest output confidence levels is smaller than an established threshold (*AmbThres*). More specifically, we also have:
  - *AC (Total):* All detected ambiguity cases.
  - *AC (Error):* Only those ambiguity cases related to error cases (observed and inferred outputs are different).
  - *EC:* Error cases. Number of remaining cases for which the observed and inferred output classes are different.
- *Data (TOTAL):* The total number of instances in the dataset.
- Error cases (*EC*): Number of remaining cases for which observed and inferred values are different.

A good KB should minimise all of them by offering an accurate (reducing EC), consistent (reducing AC) and complete (reducing UC) set of rules. They can be combined to define the next accuracy index:

$$Accuracy = 1 - \frac{EC + AC(Error) + UC}{DATA(TOTAL)}$$

$$Accuracy(CONS) = 1 - \frac{EC + AC(TOTAL) + UC}{DATA(TOTAL)}$$

$$Accuracy(BT = 0) = 1 - \frac{EC + AC(Error) + UC(Error)}{DATA(TOTAL)}$$

Table 11: Quality Measurements

KB	cv%	ac	acons	abt=0	acfd	micfd	macfd	me
FDT1	100	0.914	0.908	0.914	0.865	0	1	3
FDT4	100	0.424	0.303	0.424	0.516	0	1	3
FDT7	100	0.873	0.87	0.873	0.448	0	0.842	3
FDT15	100	0.717	0.704	0.717	0.704	0.013	1	3
FDT14	99.794	0.939	0.934	0.939	0.925	0	1	3
FDT2	100	0.935	0.928	0.935	0.701	0	1	3
FDT3	100	0.318	0.277	0.318	0.373	0.005	0.742	3
FDT5	100	0.424	0.303	0.424	0.468	0	1	3
FDT6	100	0.930	0.924	0.930	0.645	0.002	1	3
FDT9	100	0.618	0.549	0.618	0.361	0	1	3

VII. Evaluation

The evaluation was carried out over a variety of experimentations. More specifically, the suitable fuzzy partition was investigated, and subsequently we induced the corresponding decision rules and calculated and assessed the quality criteria to measure the accuracy of each approach. Table 11 clearly shows that the best results correspond to FDT1, FDT2, FDT3, FDT5, FDT6, FDT9, FDT4, FDT7, FDT14, and FDT15, which are based on different algorithms for induction and partitions, with the following parameters:

- Coverage (cv%): percentage of data samples from the selected dataset that fire at least one rule in the rule base with an activation degree higher than the pre-defined Blank threshold (BT).
- Accuracy (ac): percentage of data samples properly classified

- Accuracy (*acons*): percentage of data samples properly classified.
- Average Confidence Firing Degree (*acfd*): mean value of the firing degree related to the winner rule for the whole dataset.
- Minimum Confidence Firing Degree (*micfd*): minimum value of the firing degree related to the winner rule for the whole dataset.
- Maximum Confidence Firing Degree (*macfd*): maximum value of the firing degree related to the winner rule for the whole dataset.
- Max Error (*me*): maximum difference between the observed class and the inferred one.
- Mean Square Classification Error (*msce*)

FDT14 was generated by attributes resulting from the text mining extraction (RRs and QRS) integrated with K-means and HFP algorithms, gave the best result with a classification rate of 93.9% and a 0.646 interpretability value. On the other hand, FDT15, created by expert fuzzy partition with fuzzy decision tree induction algorithm, and it gave a lower classification rate and interpretability value of 71.7% and 0.025 respectively.

Considering FDT1 (regular fuzzy partition with fuzzy decision tree induction algorithm), FDT4 (K-means fuzzy partition and fuzzy decision tree induction algorithm), and FDT7 (regular fuzzy partition and fuzzy decision with Wang and Mendel induction algorithm), we noticed that the interpretability value was zero, this is clearly explained by the very large number of rules.

A. Analysis of Rules:

Rules

Rule	Type	Active	If RR0	AND RRs	AND COMP	AND QRS	AND PP	AND ENERGIE
1	E	yes			Late L			
2	E	yes		NOT(Irregular R)	NOT(Late L)	Small		
3	E	yes	Irregular L		NOT(Late L)	Average		
4	E	yes	NOT(Irregular L)			Average		
5	E	yes			Late R	Large	Small	
6	E	yes		Irregular R	NOT(Late L)	Small		
7	E	yes			Regular	Large	Small	
8	E	yes			NOT(Late L)	Large	Average	
9	E	yes			NOT(Late L)	Large	Tall	
10	E	yes	Irregular L	Irregular R	Late R	Large	Tall	high
11	E	yes	Irregular R	Irregular L	Late L	Average	Tall	high

Figure 10: Rules of FDT14 (expert & text mining) and fuzzy decision tree algorithm.

In this section, we further discuss the FDT14 results, which have been shown to be more accurate, with a knowledge base consisting of the "data & expert" parts as a system partition.

More specifically, we successfully built a simplified and optimised based knowledge, with 11 decision rules and 9 induced rules from the database and 2 of the expert. Figure 10 depicts an example based on one sample from the data-set, which activates both rule 8 and rule 9 with 0.553 and 0.447 distributed as Class 2.

We also had an improvement of the interpretability as shown by the corresponding index of 0.646, providing a very good compromise between the accuracy, with value of 0.939, and interpretability. Furthermore, this also yields Nauck's index [26] of 99.796, defined by the product

$$\text{Nauck}' s = \text{comp} \times \text{part} \times \text{cov}$$

where

- *Comp* represents the complexity of a classifier measured as the number of classes divided by the total number of premises.
- *Part* stands for the average normalized partition index overall input variables. It is computed as the inverse of the number of labels minus one (two is the minimum number of linguistic terms in a partition) for each input variable.
- *Cov* is the average normalized coverage degree of the fuzzy partition.

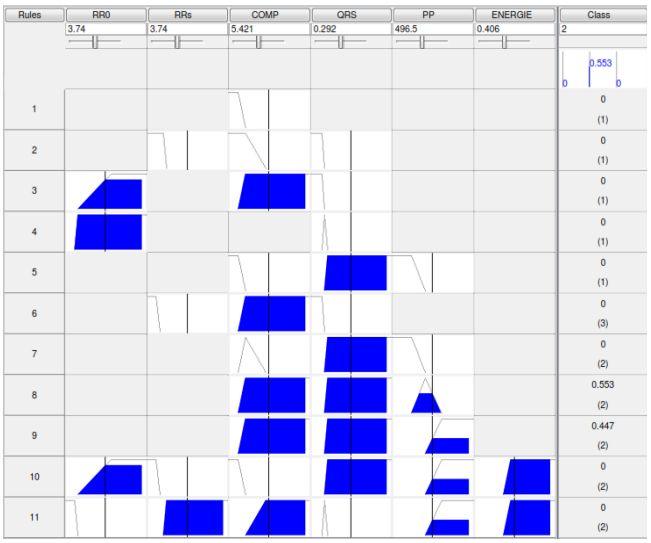

Figure 11: Inference rules

Table 12: Fingrams Measurement

INDICES	***Values***
Converge (%)	**99.796**
Accuracy	**0.939**
Average Confidence Firing Degree	**0.925**
Total Rule Length	**36**
Inferential Fired Rules (training)(Max)	**5**
Inferential Fired Rules (training)(Average)	**2.342**
Inferential Fired Rules (training)(Min)	**1**
Accumulated Rule Complexity	**11.213**
Interpretability Index (Fingrams)	**0.646**
Interpretability Index (HLIK)	**0.123**
Nauck's Index	**99.796**

VIII. CONCLUSION

In this paper we have discussed a method to build a system based on rules, using three sources of knowledge. The use of fuzzy logic, as a platform for communication between the different sources of knowledge, proves to be a successful solution to manage the fusion of knowledge in a database of common rules. The application of specific text mining methods in the extraction of knowledge from the large textual data-sets provided by PubMed, has enabled an accuracy of 93.9%, and interpretability index 0.646. This is clearly a marked improvement compared to the existing algorithms, which may obtain high accuracy but lacking in interpretability.

Furthermore, our method offers more flexibility and transparency in the system of detection, allowing expert's contribution to facilitate and guide the process of medical decision making.

REFERENCES

[1] World Health Report 2013, Retrieved 06 10, 2014, from http://www.who.int/whr/en/

[2] Alonso J.M. and Luis Magdalena L., An Experimental Study on the Interpretability of Fuzzy Systems, 2009.

[3] Gacto M. J., Alcala R., and Herrera F., Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures, Information Sciences, 181(20):4340-4360, 2011 (DOI:10.1016/j.ins.2011.02.021).

[4] Alonso J. M., Interpretable fuzzy systems modeling with cooperation between expert and induced knowledge, PhD Thesis, 2007.

[5] Zhang J., and Huang M.L. Density Approach: a New Model for BigData Analysis and Visualization, Concurrency and Computation: Practice and Experience, 2014, 1532-0634

[6] Rumelhart. D., Hinton. G & Willams. R. Learning internal representations by error propagation, in parallel distribution proceeding: exploration in the Microstructure of Cognition, foundations edited by D. Rumelhart and J. McClemlland, MIT Press, Cambridge, M.A. 1986; 1: 318-362.

[7] Meau, Y. P., Ibrahim, F., Naroinasamy, S. A. L., and Omar, R. Intelligent classification of electrocardiogram (ECG) signal using extended Kalman filter 441 (EKF) based neuro fuzzy system. Computer Methods and Programs in Biomedicine, 2006.

[8] Yu S. and Chou T., Integration of independent component analysis and neural networks for ECG beat classification, Expert Systems with Applications, Volume 34, Issue 4, 2008

[9] Hosseini H.G., Luo D. and Reynolds K.J., The comparison of different feed forward neural network architectures for ECG signal diagnosis, Medical engineering & physics 28 (4), 372-378.

[10] Raghupathi W., Data Mining in Health Care. In Healthcare Informatics: Improving Efficiency and Productivity. Edited by Kudyba S. Taylor & Francis; 2010:211-223.

[11] Casado R., and Younas M. Emerging Trends and Technologies in Big Data Processing. Concurrency and Computation: Practice and Experience 2014, 1532-0634

[12] Raghupathi W, and Raghupathi V., Big data analytics in healthcare: promise and potential, Health Information Science and Systems 2014, 2:3.

[13] Manning, C. D. Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[14] De Marneffe, M. F., MacCartney, B., and Manning, C. D. Generating Typed Dependency Parses from Phrase Structure Parses, LREC, 2006.

[15] Liu B. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012

[16] PubMed. Retrieved 06 10, 2014, from http://www.ncbi.nlm.nih.gov/pubmed/

[17] Schönbauer, R., Sommers, P., Misfeld, M., Dinov, B., Fiedler, F., Huo, Y. and Arya, A. Relevant ventricular septal defect caused by steam pop during ablation of premature ventricular contraction. Circulation, 2013

[18] Soheilykhah, S., Sheikhani, A., Sharif, A. G. and Daevaeiha, M. M. Localization of premature ventricular contraction foci in normal individuals based on multichannel electrocardiogram signals processing. Springerplus, 486, 2013.

[19] Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng in Med and Biol 20(3):45-50, 2001.

[20] Pan J and Tompkins W.J., A Real-Time QRS Detection Algorithm IEEE Transactions ON Biomedical Engineering, Vol. BME-32, NO. 3, 1985.

[21] Guillaume S. and Magdalena L., Expert guided integration of induced knowledge into a fuzzy knowledge base, Soft Computing, 2006, 773-784, Vol. 10.

[22] Piotrkiewicz M., Kudina L., Mierzejewska J., Jakubiec M. and Hausmanowa-Petrusewicz I., Age-related change in duration of after hyperpolarization of human motoneurones, The Journal of Physiology, 585, 483-490, 2007.

[23] Casillas J., Accuracy Improvements in Linguistic Fuzzy Modelling, Springer Science & Business Media, 2003

[24] Chen T., An effective fuzzy collaborative forecasting approach for predicting the job cycle time in wafer fabrication, Computers & Industrial Engineering, Volume 66, Issue 4, 2013

[25] Guillaume, S., and Charnomordic, B. Fuzzy inference systems: An integrated modeling environment for collaboration between expert knowledge and data using FisPro. Expert Systems with Applications, 8744-8755, 2012.

[26] D.D. Nauck, Measuring interpretability in rule-based classification systems, In Proceedings of the FUZZ-IEEE, St. Louis, Missouri, USA, pp. 196–201, 2003 (DOI:10.1109/FUZZ.2003.1209361).