Applied Soft Computing 13 (2013) 3439–3448

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

Fuzzy rules for describing subgroups from Influenza A virus using a multi-objective evolutionary algorithm

C.J. Carmona[a,*], C. Chrysostomou[b], H. Seker[c], M.J. del Jesus[a]

[a] Department of Computer Science, University of Jaén, 23071 Jaén, Spain
[b] Department of Genetics, University of Leicester, Leicester LE1 7RH, United Kingdom
[c] Centre for Computational Intelligent, School of Computing, De Montfort University, Leicester LEI 9BH, United Kingdom

ARTICLE INFO

Article history:
Received 18 July 2012
Received in revised form 28 January 2013
Accepted 11 April 2013
Available online 25 April 2013

Keywords:
Evolutionary fuzzy system
Influenza A virus
Neuraminidase
Subgroup discovery
Multi-objective evolutionary algorithm

ABSTRACT

Extraction of biologically-meaningful knowledge is one of the important and challenging tasks in bioinformatics, in particular computational analysis of DNA and protein sequences, in order to identify biological function(s) and behaviour(s) of newly-extracted sequences. Computational intelligence techniques in corporation with sequence-driven features have been applied to tackle the problem and help classify different functional classes of the sequences. In order to study this problem, subgroup discovery algorithms together with a signal processing-based feature extraction method are applied, where the sequences are represented as a signal. The applicability of this method has been studied through four different Neuraminidase genes of Influenza A subtypes, H1N1, H2N2, H3N2 and H5N1. The results yielded not only higher predictive accuracy over these four classes of the proteins but also interpretable rule-based representation of the descriptive model with a significantly reduced feature set driven by means of the signal processing method. Subgroup discovery technique based on evolutionary fuzzy systems is expected to open new areas of research in bioinformatics and further help identify and understand more focused therapeutic protein targets.

1. Introduction

In recent years, decoding the rules that drive biological functions of the proteins directly from the primary structure of a protein sequence has become a subject of intensive research, one example begin the Basic Local Alignment Search Tool (BLAST) [1]. Signal processing techniques are also being used in bioinformatics to extract information that is expected to match protein biological functions, examples being the Resonant Recognition Model [2–4] and Complex Resonant Recognition Model [5]. Our study is performed through subgroup discovery (SD) techniques in order to search for similarities and differences between different subtypes of Influenza A, from the proteins based on features extracted from signal processing techniques. These results can then help to understand these subtypes of proteins in the pharmacological industry.

Signal processing techniques can generate large amounts of data which can be related to protein biological functions. The Resonant Recognition Model and Complex Resonant Recognition Model is only one technique that tries to identify which of the features extracted are related to the protein biological function. A new method is needed that is able to identify all the useful features that can be related to the characterisation of Influenza A virus problem and discarding any ineffective or noisy data. For this problem, SD [6,7] may be suitable because it could determine whether a feature or a set of features extracted from protein sequences using signal processing techniques can be used to characterise different protein classes.

The SD data mining task is somewhere halfway between predictive and descriptive induction and its objective is to find interpretable rules to describe unusual relationships in the data. Among the bibliography different SD algorithms applied to bioinformatic problems can be found [8–10] (specifically in cancer environment). These problems were characterised by a high dimensionality with respect to the number of variables (between 7000 and 22,000), and a low number of examples (not more than 200 examples). In these papers, a good behaviour of the SD algorithms in solving bioinformatic problems was presented, where novelty and interpretable models were obtained.

Interpretability is a key factor for the SD algorithm where partial relations with unusual and interesting behaviour instead of complete relations are more interesting for their simplicity. To do so, SD algorithms allow to obtain knowledge that let us to better understand the problem under study. These algorithms tend to discard non-relevant information, and although the primary objective of any SD algorithm is not to predict but to describe, the obtained rules are usually very accurate so they can be used to classify. These

* Corresponding author. Tel.: +34 953 211956; fax: +34 953 212472.
E-mail address: ccarmona@ujaen.es (C.J. Carmona).

1568-4946/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2013.04.011

properties make it interesting to apply SD techniques for the classification and characterisation of influenza A virus problem. In the specialised bibliography of SD different approaches can be found as can be observed in the review [11]. However, for this problem, with a high number of features, the SD algorithms based on evolutionary fuzzy systems (EFSs) [12] have special relevance specially for the use of fuzzy logic [13] and genetic algoritms [14] which gives a high interpretability capacity and efficiency in the search process. Using EFS for SD task has shown good results for problems with large number of variables and for this reason, Influenza A virus experts consider that SD algorithms based on EFSs could provide useful information to the problem. Within the EFSs for SD, different algorithms can be found SDIGA [15], MESDIF [16] and NMEEF-SD [17], which have been applied in different real-world applications like medicine [18] or e-learning [19] showing their good behaviour. In addition, due to the nature of the problem, experts in the analysis of the Influenza A virus consider that EFSs could provide new information to the problem.

In this paper, an exhaustive experimental study with the NMEEF-SD algorithm is presented for the problem due to the remaining EFSs for SD obtain results with low quality. The study is tackled from two different perspectives. On the one hand, a complete study from the point of view of the SD task is performed to find the best parameters employed for the algorithm to solve the problem with respect to interpretability, novelty and precision. On the other hand, a predictive analysis is performed to show the ability of the NMEEF-SD algorithm to classify new proteins in the different virus studied. Finally, the model obtained by the NMEEF-SD with the complete data set is presented in order to show the more representative subgroups for the Influenza A virus problem.

The paper is organised as follows: Section 2 presents the Influenza A Virus protein sequences, Section 3 describes the signal processing methods used to extract protein related features, Section 4 describes the SD problem, Section 5 presents the characteristics of the SD algorithm used and the main contributions of this algorithm to the Influenza A virus problem, Section 6 shows the experimental study which examines the problem from two perspectives: descriptive and predictive. Finally, some concluding remarks are outlined in Section 7.

2. Protein sequences

Influenza A virus belongs to the Orthomyxoviridae family of viruses and can affect mainly birds and some mammals. The Influenza A virus genome consist of eight single genes; the hemagglutinin (HA) gene, the neuraminidase (NA) gene, the nucleoprotein (NP) gene, the matrix proteins (M) gene, the non-structural proteins (NS) gene and three RNA polymerase (PA, PB1, and PB2) genes. Human pandemic outbreaks sometimes occur when Influenza A virus' are transmitted from wild birds to domestic poultry. During the twentieth century three major Influenza A pandemics were recorded, caused by H1N1, H2N2, and H3N2 viruses. In addition H5N1 virus is considered as a current pandemic threat. For this analysis four different subtypes of Influenza A virus Neuraminidase gene were used as this is the target for current antiviral drugs, called neuraminidase inhibitors [20]. For Influenza A subtypes 200 H1N1 NA proteins from 2009, 76 H2N2 NA proteins from the period 1957–1968, 200 H3N2 NA proteins from the period 1968–2000 and 70 H5N1 NA proteins from the period 2005–2009 were collected from the Influenza Virus Resource data set [21]. The relationship of Influenza A subtypes with respect of the NA gene is the following:

- H1N1 from 2009 is the result of reassortment between the Eurasian H1N1 Influenza A swine virus and the H1N2 swine virus [22]. H1N1 retains the NA gene from the Eurasian H1N1 Influenza A swine virus.
- H2N2 from the period 1957–1968 is the result of reassortment between existing human H1N1 and avian H2N2 viruses [22]. H2N2 retains the NA gene from the avian H2N2 virus.
- H3N2 from the period 1968–2000 is the result of reassortment between circulating human H2N2 and avian H3 viruses [22]. H3N2 retains the NA gene from the human H2N2 virus.
- H5N1 from the period 2005–2009 was created by combining various Influenza A subtype virus' [23], where H5N1 retains the NA gene from the avian H1N1 virus.

Table 1
Average percent identity.

	H1N1	H2N2	H3N2	H5N1
H1N1	93%	–	–	–
H2N2	42%	96%	–	–
H3N2	40%	86%	94%	–
H5N1	83%	43%	41%	96%

Percentage identity is a measurement used to determine the similarity between protein sequences. By using CLUSTALW, a freely available online tool [22], pairwise percent identity of all the subtype Influenza NA genes was calculated. Table 1 shows the average percentage identity between all the classes.

As Table 1 shows, the percent identity within each individual Influenza subtype class is very high with 93%, 96%, 94% and 96% for H1N1 NA, H2N2 NA, H3N2 NA and H5N1 NA Influenza A subtypes. In contrast to the individual class, percent identity from different classes may vary significantly with high average percent identity of 83% between H1N1 and H5N1 and 86% between H2N2. Very low average percent identity was determined between H1N1 and H2N2 with 42%, H1N1 and H3N2 with 40%, H5N1 and H2N2 with 43%, and finally H5N1 and H3N2 with 41% average percent identity.

3. Signal processing for protein sequence analysis

Using digital signal processing techniques the goal is to extract information that can be related to biological functions of proteins. Various methods have been used in bioinformatics for analysing protein sequences in recent years where one of the most common methods is the Resonant Recognition Model [2–4] and Complex Resonant Recognition Model [5]. Previous studies [24] used Influenza A subtypes to analyse the HA gene with the resonant recognition model aiming to identify new therapeutic targets for drug development by better understanding the interaction of the Influenza virus and its receptors.

In contrast to previous studies, the analysis was performed directly on the absolute spectrum which derives by applying Discrete Fourier Transform (DFT) to each numerically encoded protein sequence. Electron-ion interaction potential (EIIP) [25,26] amino

Table 2
EIIP values.

Amino acid	EIIP	Amino acid	EIIP
Leu	0.0000	Tyr	0.0516
Ile	0.0000	Trp	0.0548
Asn	0.0036	Gln	0.0761
Gly	0.0050	Met	0.0823
Glu	0.0057	Ser	0.0829
Val	0.0058	Cys	0.0829
Pro	0.0198	Thr	0.0941
His	0.0242	Phe	0.0946
Lys	0.0371	Arg	0.0959
Ala	0.0373	Asp	0.1263

acid index as shown in Table 2 is used to express protein sequences in order to numerical sequences to be able to apply DFT.

3.1. Preprocessing the protein sequences

By applying pre-processing techniques to the signals such as zero-padding and windowing studies have shown that the features extracted from signal processing techniques can be influenced [27]. Before applying DFT to the protein sequences, zero-padding and windowing, used in signal processing needs to be considered.

The first technique under investigation is the windowing where a pre-calculated window is multiplied to the encoded numerical sequences in order to reduce spectral leakage. For this study, the Hamming window [28] is selected, and can be computed using Eq. (1).

$$w = 0.54 - 0.46\cos\left(\frac{2\pi(N-1)}{N-1}\right) \tag{1}$$

The second technique under investigation is the zero-padding [29,30] where to order to increase signal length a number of zero elements are supplemented to the end of individual sequence. This practice is necessary as the given protein sequences may not have the same length.

3.2. Discrete Fourier transform

The Discrete Fourier transform (DFT) is defined as follows:

$$X(n) = \sum_{m=0}^{N-1} x(m)e^{-j(2/N)nm}, \quad n = 1, 2, \ldots, N/2 \tag{2}$$

where $x(m)$ is the mth member of the numerical series, N is the total number of points in the series, and $X(n)$ are coefficients of the DFT. The following formula determines the maximal frequency in the spectrum:

$$F = \frac{1}{2d} \tag{3}$$

where F is the maximal frequency of all signals and d is the distance between points of the sequence.

If it is assumed that all points of the sequence are equidistant with distance $d = 1$ then the maximum frequency in the spectrum can be found as $F = 1/2(1) = 0.5$. This indicates that the frequency range does not depend on the number of points in the sequence but only on the resolution of the spectrum. The output of DFT is a complex sequence and can be represented as follows:

$$X(n) = (R(n) + I(n)j), \quad n = 1, 2, \ldots, N/2 \tag{4}$$

where $R(n)$ is the real part of the sequence and $I(n)j$ the Imaginary part.

The final step is calculating absolute spectrum from DFT complex sequence, which can be formulated as follows:

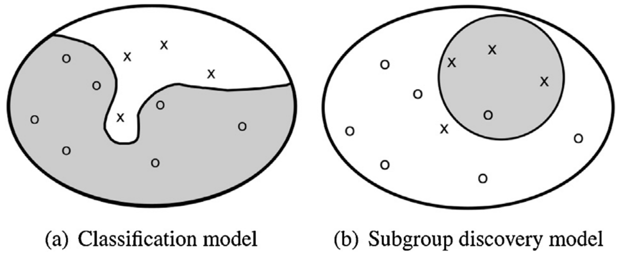
$$S_a(n) = X(n)X*(n) = \left|X(n)\right|^2, \quad n = 1, 2, \ldots, N/2 \tag{5}$$

where S_a is the absolute spectrum for a specific protein, $X(n)$ are the DFT coefficients of the series $x(n)$ and $X^*(n)$ are the complex conjugate. Next Eq. (6) is used to scale absolute spectrum:

$$V = \frac{\sqrt{\sum_{n=0}^{L} C_a(n)}}{L} \tag{6}$$

where L is the number of points in the Absolute (S_a) spectrum.

For the analysis of Influenza A virus proteins, as the sequences have different lengths zero-padding was used to extend all protein sequences to $N = 512$; thus the output of absolute spectrum (Eq. (5)) is 256 features.



(a) Classification model

(b) Subgroup discovery model

Fig. 1. Models obtained with different data mining approaches for the same data set.

4. Subgroup discovery

SD [6,7] is a broadly applicable data mining task aimed at discovering interesting and relevant relationships between properties of a data set with respect to a specific property which is of interest to the user.

The following subsections present the main elements and properties of SD: Section 4.1 presents an introduction to SD task and Section 4.2 shows the utility of the EFSs in solving the SD problem.

4.1. Introduction to the subgroup discovery task

The concept of SD was initially introduced by Kloesgen [6] and Wrobel [7]. It can be defined in this way [31]:

> "In subgroup discovery, we assume we are given a so-called population of individuals (objects, customer, . . .) and a property of those individuals we are interested in. The task of subgroup discovery is then to discover the subgroups of the population that are statistically "most interesting", i.e. are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest."

Hence the main property is the search for relationships between a group of variables with respect to a target variable, where the objective is the extraction of partial relationships with unusual and interest behaviour instead of complete relations.

An induced subgroup can be represented by a rule (R):

$$R : Cond \quad \rightarrow \quad Class$$

where *Class* is a value for the variable of interest for the SD task, and *Cond* is commonly a conjunction of features (attribute-value pairs) which is able to describe an unusual statistical distribution with respect to the *Class*.

SD is somewhere halfway between predictive and descriptive induction and it is differentiated from classification techniques basically because SD attempts to describe knowledge for the data while a classifier attempts to predict it. Furthermore, the model obtained by a SD algorithm is usually simple and interpretable, while that obtained by a classifier is usually more complex and precise.

As can be observed in Fig. 1 the model obtained by the classifier – Fig. 1(a) – is more complex than the model obtained by the SD approach – Fig. 1(b) –. In addition, the accuracy of the classification model is greater than the accuracy obtained by the SD model, but with respect to the interpretability the best results are obtained by the SD model. In conclusion, a SD algorithm obtains simple models for describing unusual and significant behaviour of the data with a good level of accuracy.

The election of quality measures is one of the most important elements in order to apply a SD algorithm [32]. Throughout the literature there was an absence of consensus with respect to the use of quality measures in SD, for example in [6,33–35] authors used different quality measures in order to analyse SD algorithms. However, in [11] an overview of the quality measures used throughout the literature are presented with respect to the characterisation of them, i.e. considering definitions presented and properties desired by SD algorithms, an approach must satisfy: interpretability, novelty and good relation between sensitivity-confidence. Therefore, in this paper the following quality measures are used in order to analyse quality of the algorithms employed:

- *Sensitivity* [6] is used to quantify the quality of individual rules according to the individual patterns of interest covered. This is a measure with characteristics of precision and generality and it can be computed as:

$$Sens(R) = \frac{tp}{tp+tn} = \frac{tp}{Pos} \tag{7}$$

  where *tp* are the examples correctly classified, and *Pos* are the examples of the *Class* specified by the rule.
- *Unusualness* [36], which attempts to obtain a tradeoff between generality, interest and precision in the results, and can be computed as:

$$Unus(R) = \frac{tp+fp}{n_s} \cdot \left(\frac{tp}{tp+fp} - \frac{Pos}{n_s}\right) \tag{8}$$

  where *fp* are the examples incorrectly classified, and n_s are the examples of the data set.
- *Confidence* [37] measures the relative frequency of examples satisfying the complete rule among those satisfying only the antecedent. In this paper one modified confidence measure for fuzzy rules, *Fuzzy Confidence* [15] is used, which computes the confidence for fuzzy subgroups. It is defined as:

$$FCnf(R) = \frac{\sum_{E^k \in E/E^k \in Class} APC(E^k, R)}{\sum_{E^k \in E} APC(E^k, R)} \tag{9}$$

  where *APC* (Antecedent Part Compatibility) is the degree of compatibility between an example and the antecedent part of a fuzzy rule, i.e. the degree of membership for the example to the fuzzy subspace delimited by the antecedent part of the rule and E^k is a set of examples, where *Class* is the value of the target variable for the example E^k (i.e. the class for this example).
- *Significance* [6] which indicates the significance of a finding, if measured by the likelihood ratio of a rule, and it can be computed as:

$$Sign(R) = 2 \cdot \sum_{k=1}^{n_c} tp_k \cdot log \frac{tp_k}{Pos_k \cdot (tp_k + fp_k)/(n_s)} \tag{10}$$

  where n_c is the number of *Classes* to study. It must be noted that although each rule is for a specific *Class*, the significance measures the novelty in the distribution impartially, i.e. significance of a rule is calculated for all values of the *Class* although rule is obtained for one value of the *Class*.

With this group of measures and the study of interpretability of the subgroups obtained the quality of a SD model can be calculated and analysed in a correct and quantifiable way.

4.2. The use of evolutionary fuzzy systems to find unusual relations in the characterisation of Influenza A virus

Different studies aiming to solve real problems using classical SD approaches in the bioinformatic domain [8–10] have been presented throughout the literature. More specifically, the main objective of these studies were to obtain gene expression subgroups in order to detect diagnoses related to the cancer. Bioinformatic domains are characterised by the high dimensionality of the data set where a wide number of variables and a low number of instances can usually be found. This high dimensionality characteristic is a problem when we are trying to obtain accurate and precise models for the algorithms, and in general the models obtained have a high number of rules and variables, making it difficult to describe the data correctly. In this case, the SD algorithms obtain simple rules with a high level of precision and accuracy. This property makes them very suitable for solving the characterisation of Influenza A virus. A large number of algorithms can be observed in the literature for SD: algorithms based on classification approaches, algorithms based on association approaches and evolutionary algorithms. However, in this paper EFSs are used because:

- All variables are real and in this situation the fuzzy sets lets us represent knowledge in an intuitive way and closer to human reasoning, which makes knowledge extracted and represented in the rules – in this case, fuzzy rules – more interpretable.
- The huge dimensionality of the data is a problem for algorithms such as Apriori-SD or CN2-SD, and it is impossible to obtain results with them. SD algorithsm based on EFSs can handle these kind of problem properly.

EFSs are basically fuzzy systems augmented by a learning process based on evolutionary computation [12], which includes genetic algorithms, genetic programming and evolutionary strategies, among other evolutionary algorithms [38].

Fuzzy systems are one of the most important areas for the application of the fuzzy set theory [13]. Usually these systems consider a model structure in the form of fuzzy rules. The use of these systems in the algorithms avoids the need to perform a previous discretisation to analyse the data, because this previous step could lead to a loss of information in the model obtained. Furthermore, the interpretability of the rules is improved because the experts can study the behaviour of different properties of the problem with linguistic labels such as *Low*, *Normal* or *High*, depending on the definition of the problem, instead of numbers or intervals. Within the specialised literature a wide number of papers related to the use of fuzzy logic can be observed such as [39–41].

On the other hand, evolutionary algorithms [42] imitate the principles of natural evolution in order to form searching processes. The most widely used are the genetic algorithms which are inspired by natural evolution processes [43]. This type of algorithms has solved different problems in real domains, for instance in satisfiability problems [44], in multi-depot homogenous locomotive assignment with time windows [45], in problems of minimum energy broadcast in wireless ad hoc networks [46] or for improving prediction accuracy in gene classification [47], among others.

As we mentioned previously, there are different EFSs for solving SD task: SDIGA [15], MESDIF [16] and NMEEF-SD [17]. Despite the good behaviour of this group of algorithms in resolving the SD problem, in this paper the NMEEF-SD algorithm is employed because

- it is a novel approach which obtains significant and accurate results, and
- MESDIF and SDIGA obtain results with low quality in previous analysis performed in this problem.

In the following section the main properties of the NMEEF-SD algorithm are shown.

5. NMEEF-SD: Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery

NMEEF-SD [17] is an EFS whose objective is to extract descriptive fuzzy and/or crisp rules for the SD task, depending on

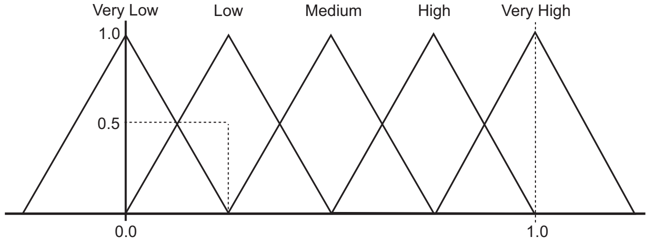


Fig. 2. Example of fuzzy partition for a continuous variable with five labels.

the type of variables present in the problem. It is based on a multi-objective approach, the NSGA-II [48] algorithm, which is a computationally fast multi-objective evolutionary algorithm based on a non-dominated sorting approach, and also on the use of elitism.

Below in Section 5.1 the most important properties of the algorithm are detailed, and finally in Section 5.2 the contribution of NMEEF-SD to the problem of the Influenza A virus Subtype protein can be observed.

5.1. Main properties of the algorithm

In NMEEF-SD algorithm fuzzy logic is used to represent the continuous variables. This use of linguistic variables allows us in data mining processes to use numerical features without the need to increase the interpretability of the extracted knowledge through discretisation. The continuous variables are considered linguistic, and the fuzzy sets corresponding to the linguistic labels can be specified by the user or defined by means of a uniform partition, if the expert knowledge is not available. Due to the absence of knowledge related to the discretisation provided by the experts, uniform partitions with triangular membership functions are used, as shown in Fig. 2 for a variable with five linguistic labels such as {*Very Low*, *Low*, *Medium*, *High*, *Very High*}, which could be represented as $X_m : \{LL_m^1, LL_m^2, LL_m^3\, LL_m^4\, LL_m^5\}$, too. In this paper, NMEEF-SD has been employed with different number of linguistic labels in order to study and analyse the results obtained.

With respect to the representation of the rule, NMEEF-SD employs the "*Chromosome = Rule*" approach, where only the antecedent is represented in the chromosome and the consequent is prefixed to one of the possible values of the target feature in the evolution. The algorithm must therefore be executed as many times as the number of different values the target variable contains. It uses an integer representation model with as many genes as variables contained in the original data set without considering the target variable. Thus the set of possible values for the categorical features is that indicated by the problem, and for numerical variables it is the set of linguistic terms determined heuristically or with expert information.

Therefore, a fuzzy rule describing a subgroup is represented as:

$$R : \mathit{If}\ X_1\ \mathit{is}\ \mathit{Low}_1\ \mathit{and}\ X_7\ \mathit{is}\ \mathit{Medium}_7\ \mathit{then}\ \mathit{Class}_j$$

considering the following:

- $\{X_m/m = 1, \ldots, n_v\}$ is a set of features used to describe the subgroups, where n_v is the number of features. These variables can be categorical or numerical.
- $\{Class_j/j = 1, \ldots, n_c\}$ is a set of values for the target variable, where n_c is the number of values.

In the multi-objective approach, the quality measures considered as objectives in the evolutionary process are selected depending on the nature of the problem to solve. In this paper, NMEEF-SD uses *Sensitivity* (Eq. (7)) and *Unusualness* (Eq. (8)) as objectives. These quality measures are employed because both measures have precision and generality features and *Unusualness* also provides interest to the process of rules extraction (these properties are analysed in the review [11]). With respect to the set of rules extracted, NMEEF-SD returns Pareto filtered with respect to a minimum confidence threshold (defined as parameter), i.e. NMEEF-SD obtains a Pareto for a value of the *Class* where the individuals (rules) are filtered through a value of confidence. With the application of this filter, only rules with high confidence are considered. This set of rules can be provided to the expert to describe and characterise this vale for the class. In the Influenza A virus problem the objective is to obtain rules to describe and predict all the class values. For this reason the NMEEF-SD algorithm is executed for all the class values and the final rule set (and so classification system) is the union of fuzzy rule set obtained for each class value (that is, the fuzzy rules in Pareto which exceed the confidence threshold).

```
Generate the initial (parents) population
while (numberofevaluations) is not reached do
    Generate offspring population
    Join the parent and offspring population in a combined one
    Generate all non-dominated fronts of the combined population
    if Pareto front (Front 1) evolves then
        Apply NSGA-II evolution
    else
        Apply Re-initialisation Based On Coverage
    end if
end while
Return the individuals of the Pareto front which reach a fuzzy
confidence threshold
```

Fig. 3. Operation scheme of NMEEF-SD algorithm.

A single operation scheme of the NMEEF-SD algorithm can be observed in Fig. 3. A complete description for the algorithm can be found in [17].

5.2. Contribution of the NMEEF-SD algorithm to the Influenza A virus problem

The main drawback of the classifiers in solving the bioinformatic problem is in general the lack of interpretability for the models obtained, because these models are extracted with accuracy as the main objective and the majority of situations are complex and use a wide number of variables to describe the different classes of the data set. In this way it is very difficult for the experts to analyse and understand the behaviour of the different classes studied.

However the SD algorithms extract very simple and interpretable models where only some rules with a low number of variables for each class are obtained. The use of the NMEEF-SD algorithm in this problem also facilitates the analysis to the experts because it uses linguistic labels in all the variables of the data set.

The search for unusual and interesting rules for the SD algorithms is another advantage provided by the NMEEF-SD algorithm. The use of unusualness (Eq. (8)) and sensitivity (Eq. (7)) as objective vectors in the multi-objective approach also provides a maximisation, not only for these measures but also for other measures in SD such as significance and confidence, because unusualness and sensitivity have precision, novelty and generality properties in their definitions. Therefore, NMEEF-SD contributes to the extraction of novelty and significance as well as knowledge about relationships between the properties of the problem and different types of the Influenza A virus.

Finally, the NMEEF-SD algorithm can be also studied as a classifier to see the behaviour of the algorithm in predicting new protein sequences introduced into the data set. Thus the experts

Table 3
Parameters for the NMEEF-SD algorithm.

Parameters employed by the NMEEF-SD algorithm
Population size=50, Evalutions = 10000, Crossover probability=0.60, Linguistic labels = 3, 5, 7 and 9, Mutation probability=0.1, Re-initialisation based on coverage (50% of biased), Minimum confidence = 0.2, 0.4 and 0.6, and Representation of the rule = Canonical

can distinguish different groups of proteins with simple and single rules, obtained to enable the experts to evaluate the data through an exhaustive description of the problem.

6. Experimental study

The main objective of this paper is to find unusual and interesting relationships among different proteins in the Influenza A virus problem with respect to the different classes of these virus. Within these relationships different similarities/differences between several subtypes of the Influenza A virus can give the experts information to understand these subtypes of virus. The problem is analysed and studied with a SD approach, the NMEEF-SD algorithm.

As mentioned above, the problem has a high dimensionality and is composed of 256 features and 546 proteins sequences, where the proteins are distributed in the classes with 200 for class H1N1, 76 for H2N2, 200 for H3N2 and 70 for class H5N1.

All features used have a real domain and are therefore continuous features, i.e. 256 continuous variables. The NMEEF-SD algorithm considers the continuous variables as linguistic fuzzy variables with fuzzy logic. More specifically, as mentioned above, in this paper uniform partitions with triangular membership functions are used.

The parameters employed by the NMEEF-SD are presented in Table 3.

Due to the non-deterministic nature of the NMEEF-SD, the algorithm is executed five times for each data set with a 5-fold cross validation. In this way, the results shown are the average of the results obtained for each data set for the different executions, i.e. the average of the 25 executions. Therefore, the following average results in the experimental study in the tables can be observed: the number of linguistic labels employed, the minimum confidence threshold used (Min_{Cnf}), number of rules (♯*Rules*), number of variables (♯*Vars*), significance (*SIGN*), unusualness (*UNUS*), sensitivity (*SENS*) and confidence (*CONF*).

The experimental study is divided into two subsections: First, in Section 6.1 the results obtained by the NMEEF-SD algorithm are studied from several points of view: SD analysis and predictive analysis. Finally, in Section 6.2 a descriptive analysis is performed for the complete data set with respect to the different classes studied in the problem, incorporating an exhaustive analysis of the subgroups extracted.

6.1. Analysis of the results obtained by the NMEEF-SD algorithm

Due to the complexity of the problem and the absence of knowledge by experts about the discretisation for the features in this problem, it is necessary to use different numbers of linguistic labels and minimum confidence thresholds in order to find the configuration of the algorithm which obtains the best results. Therefore in this experimental study 3, 5, 7 and 9 linguistic labels are studied with different minimum confidence thresholds for each one (0.2, 0.4 and 0.6).

In this way the NMEEF-SD algorithm is executed 25 times for each combination of parameters, and the average is shown for each row in Table 4. The best results for each quality measure are highlighted.

In Table 4 the best results are obtained with the use of 3 linguistic labels and more specifically with the use of a minimum confidence threshold of 0.6, as can be observed. However, the number of rules obtained is lower than the number of classes analysed in the data set, which indicates that there is some class without rules. An analysis of the subgroups extracted by the algorithm for each class with 3 linguistic labels is presented in Table 5, where all rules extracted in the cross validation are represented. In this way is tested the obtaining of rules for all values of the class.

As mentioned previously in the analysis of Table 4 and with the results shown in Table 5, the number of subgroups obtained for a minimum confidence threshold of 0.6 indicates that there are not enough subgroups to describe all the classes. This is because the confidence threshold is too high to obtain good results in all the classes. Therefore, the results obtained in this configuration must be discarded.

In summary, the best results obtained for the NMEEF-SD algorithm are obtained with 3 linguistic labels and minimum confidence of 0.2 and 0.4. To complete this statement, an analysis related to the SD task for each class in these configurations is presented below:

- The subgroups obtained for *Class H*1*N*1 have a high interpretability because the number of variables is low; in general the subgroups obtained have less than 3 variables (considering class as a variable too). The values for significance and unusualness are the highest with respect to the values obtained in the remaining class. Furthermore, the relationship between sensitivity and confidence is very good because the algorithm obtains subgroups where all the protein sequences for the class are covered and the confidence is close to 85%.

Table 4
Results obtained for the NMEEF-SD algorithm in the experimental study for the Influenza A virus problem.

LLs	Min_{Cnf}	♯*Rules*	♯*Vars*	*SIGN*	*UNUS*	*SENS*	*CONF*
3	0.2	4.60	2.79	57.945	0.153	**1.000**	0.747
	0.4	3.80	2.65	61.653	0.174	**1.000**	0.811
	0.6	2.60	2.73	**66.967**	**0.190**	**1.000**	0.849
5	0.2	3.40	2.13	47.628	0.125	0.990	0.708
	0.4	3.00	2.17	50.925	0.134	0.992	0.767
	0.6	2.20	2.10	54.155	0.148	**1.000**	0.807
7	0.2	3.00	2.28	47.832	0.110	0.963	0.760
	0.4	2.40	2.42	47.094	0.113	0.939	0.854
	0.6	1.60	2.37	52.038	0.127	0.938	**0.911**
9	0.2	1.60	2.00	40.257	0.092	0.952	0.585
	0.4	1.40	2.00	39.211	0.099	0.944	0.631
	0.6	0.60	0.80	17.191	0.048	0.378	0.394

Table 5
Results obtained for the NMEEF-SD algorithm for each class in the experimental study for the Influenza A virus problem with 3 linguistic labels.

Min_{Cnf}	*Class*	*♯Rules*	*♯Vars*	*SIGN*	*UNUS*	*SENS*	*CONF*
0.2	H1N1	8.00	2.88	69.868	0.199	1.000	0.849
	H2N2	5.00	3.20	44.562	0.101	1.000	0.543
	H3N2	6.00	2.50	64.036	0.178	1.000	0.812
	H5N1	5.00	2.60	44.907	0.102	1.000	0.717
0.4	H1N1	8.00	2.88	69.868	0.199	1.000	0.849
	H2N2	3.00	2.33	41.860	0.107	1.000	0.601
	H3N2	5.00	2.40	67.831	0.193	1.000	0.835
	H5N1	3.00	3.00	45.190	0.104	1.000	0.768
0.6	H1N1	7.00	3.00	70.349	0.202	1.000	0.867
	H2N2	0.00	0.00	0.000	0.000	0.000	0.000
	H3N2	5.00	2.40	67.831	0.193	1.000	0.835
	H5N1	1.00	3.00	44.923	0.101	1.000	0.867

Table 6
Predictive results obtained by the NMEEF-SD algorithm with 3 linguistic labels and a minimum confidence of 0.2 for the Influenza A virus problem.

Class	H1N1	H2N2	H3N2	H5N1
H1N1	0.975 ± 0.055	0.000 ± 0.000	0.000 ± 0.000	0.025 ± 0.055
H2N2	0.000 ± 0.000	0.799 ± 0.413	0.201 ± 0.413	0.000 ± 0.000
H3N2	0.000 ± 0.000	0.132 ± 0.086	0.868 ± 0.086	0.000 ± 0.000
H5N1	0.271 ± 0.424	0.000 ± 0.000	0.000 ± 0.000	0.729 ± 0.424

- For *Class H2N2* the subgroups with the lowest number of variables are obtained, so the interpretability is excellent. The values of significance and unusualness are also high considering that this class has a low number of protein sequences. The level of sensitivity obtained by the subgroups extracted is the maximum and the confidence value is good because the subgroups exceed 60%.
- In *Class H3N2* the best subgroups are obtained together with *Class H1N1s*, where the interpretability and the values of significance, unusualness, sensitivity and confidence are very high.
- *Class H5N1* is the class with the lowest number of protein sequences. In spite of this problem, the results of sensitivity and confidence are very interesting because the subgroups cover the total examples of the class with a good level of confidence (more than 70%). The results for the significance and unusualness are also very high.

Despite the fact that the objective of the NMEEF-SD algorithm is to obtain general and unusual rules to describe interesting relationships between the properties of the proteins with respect to different types of virus, the algorithm also has good behaviour as a classifier, as can be observed in the following analysis.

Table 6 shows the confusion matrix for the accuracy of the model extracted by the NMEEF-SD with three linguistic labels and a minimum confidence threshold of 0.2, and Table 7 shows the confusion matrix of the model with the same linguistic labels and 0.4 of minimum confidence. The results presented in both tables are the average of the 5-fold cross validation and the standard deviation for each one.

The total accuracy for the complete data set is of 0.872 ± 0.051 for Table 6, and 0.797 ± 0.069 for Table 7. As can be observed in this study, the model extracted by the NMEEF-SD algorithm obtains good precision for classifying new examples, although the objective of the algorithm is not to obtain a classifier but rather a set of fuzzy rules which describe knowledge about the problem and where the configuration of the algorithm with 0.2 minimum confidence threshold obtains the best results. In conclusion, NMEEF-SD shows the good behaviour of the SD algorithms in searching for unusual and novel relationships in real world applications and their excellence as classifiers, more specifically in relation to the Influenza A virus. Moreover, the behaviour shown for the algorithm gives the experts suitable information to study this virus from other points of view. The main property of the algorithm is a high interpretability (low number of variables used among the total number) which facilitates the analysis. A specific analysis for each subtype of virus can be observed below:

- For H1N1 subtype class the average accuracy is 0.965±0.055 where the misclassified proteins were the same as H5N1.
- For H2N2 subtype class the average accuracy is 0.799±0.413 where the misclassified proteins were the same as H3N2.
- For H3N2 subtype class the average accuracy is 0.868±0.086 where the misclassified proteins were the same as H2N2.
- For H5N1 subtype class the average accuracy is 0.729±0.424 where the misclassified proteins were the same as H1N1.

This analysis shows a strong correlation between the features extracted from the protein sequences using absolute spectrum and protein percentage identity between classes, as shown in Table 1. Only subtype classes that present high percentage identity between them as H1N1 with H5N1 subtype (83%) and H2N2 with H3N2 subtype (86%) were partially misclassified.

Table 7
Predictive results obtained by the NMEEF-SD algorithm with 3 linguistic labels and a minimum confidence of 0.4 for the Influenza A virus problem.

Class	H1N1	H2N2	H3N2	H5N1
H1N1	0.975 ± 0.056	0.000 ± 0.000	0.000 ± 0.000	0.025 ± 0.056
H2N2	0.107 ± 0.174	0.413 ± 0.537	0.481 ± 0.457	0.000 ± 0.000
H3N2	0.032 ± 0.025	0.043 ± 0.112	0.925 ± 0.106	0.000 ± 0.000
H5N1	0.629 ± 0.458	0.000 ± 0.000	0.029 ± 0.064	0.343 ± 0.480

Table 8
Subgroups obtained for the NMEEF-SD algorithm for each class. Results associated with each subgroup in the complete data set.

Subgroup	*SIGN*	*UNUS*	*SENS*	*CONF*
IF (f_{44} = Low AND f_{97} = Low) THEN Cl = H1N1	363.485	0.224	1.000	0.966
IF (f_9 = Low AND f54 = Low AND f_{153} = Low AND f_{217} = Low) THEN Cl = H2N2	227.960	0.105	1.000	0.600
IF (f_8 = Low) THEN Cl = H3N2	373.894	0.182	1.000	0.730
IF (f_{141} = Low AND f_{207} = Low AND f_{219} = Low) THEN Cl = H3N2	309.357	0.196	0.995	0.966
IF (f_{115} = Low) THEN Cl = H5N1	188.813	0.097	1.000	0.677

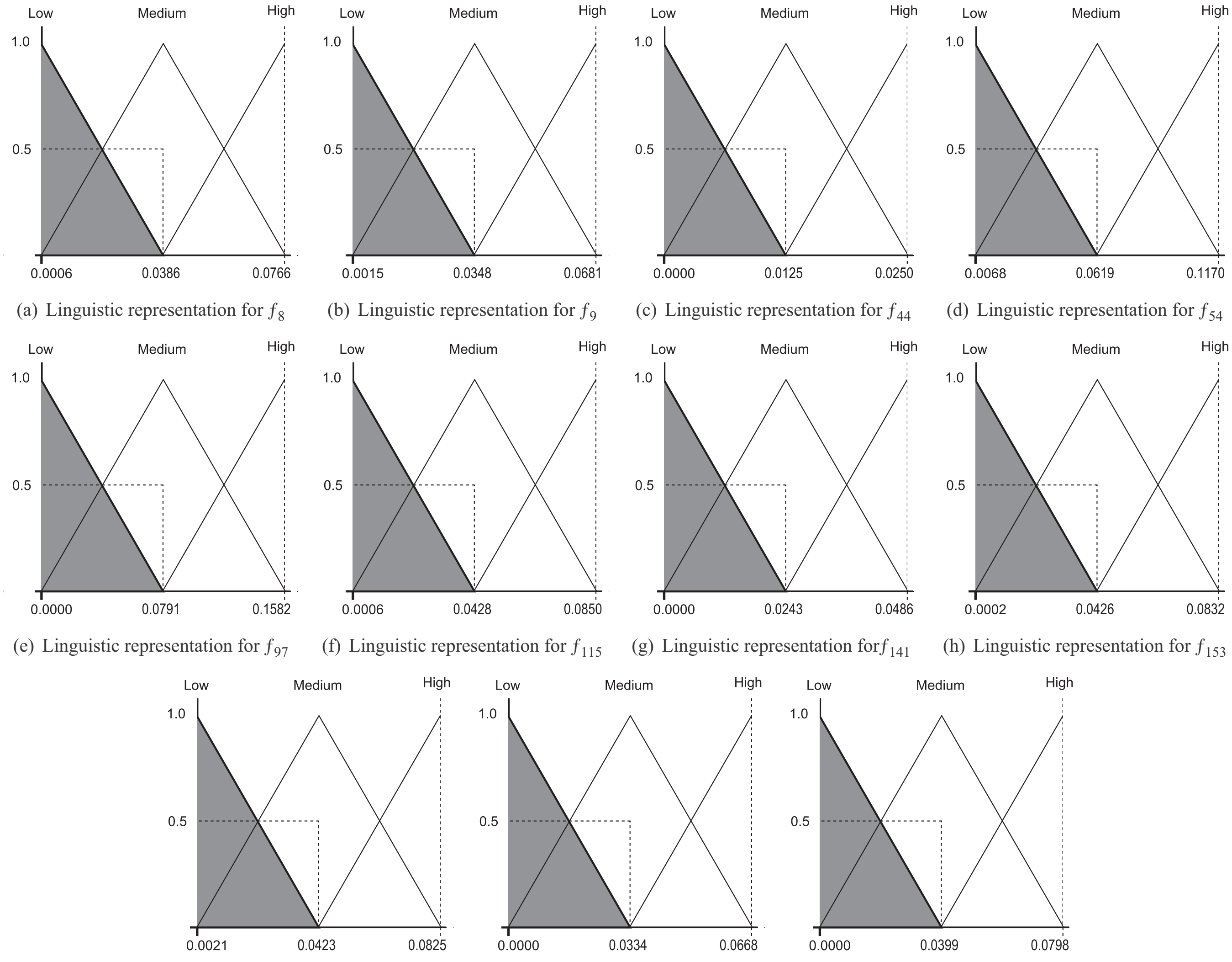
6.2. Fuzzy subgroups extracted by the NMEEF-SD

Once determined that NMEEF-SD with 3 linguistic labels and a minimum confidence threshold of 0.2 obtains the best results for the Influenza A virus problem, a new experiment was performed using the complete data set in order to analyse the subgroups obtained by the NMEEF-SD.

Table 8 shows the subgroups obtained for the NMEEF-SD algorithm for each class with 3 linguistic labels and a minimum confidence of 0.2, where the variable f_x corresponds to the feature number x. In addition, the table presents the results for each subgroup.

As can be observed in Table 8 the good results of unusualness and significance show the innovation brought by these subgroups to the problem. Furthermore, the sensitivity obtained for the majority of the subgroups is the maximum level and the confidence is very high with values higher than 0.600 and some very close to the maximum level. These good relations between the values of sensitivity and confidence represent subgroups of high quality. In addition, the interpretability of these rules is excellent with subgroups which in any case do not exceed four features.

Other methods that use signal processing techniques to extract biologically related features in order to characterise protein sequences like the Resonant Recognition Model in the HA gene [24] and the Complex Resonant Recognition for the NA gene [5] use informational spectrum analysis to retrieve these features. The extracted features are then used to characterise a specific class or compare it with another protein class based on common frequency



(a) Linguistic representation for f_8

(b) Linguistic representation for f_9

(c) Linguistic representation for f_{44}

(d) Linguistic representation for f_{54}

(e) Linguistic representation for f_{97}

(f) Linguistic representation for f_{115}

(g) Linguistic representation forf_{141}

(h) Linguistic representation for f_{153}

(i) Linguistic representation for f_{207}

(j) Linguistic representation for f_{217}

(k) Linguistic representation for f_{219}

Fig. 4. Linguistic representations of the continuous feature of the model extracted by the NMEEF-SD algorithm

peak [5]. By using the NMEEF-SD algorithm simple rules, as Table 8 shows, can be extracted based on the features retrieved from an absolute spectrum. By using these features new knowledge can be extracted and associated to the Influenza A proteins sequences. These rules created on the basis of the features extracted can then help in the understanding and development of therapies. For the Influenza A problem the rules created are based on 11 features of the absolute spectrum for all subtype classes. For example, for an unknown protein sequence to be able to determine in which subgroup it belongs only 11 features of the absolute spectrum need to be considered and not the whole spectrum, where for the Influenza A problem it consists of 256 variables. The importance of this outcome is that by using the NMEEF-SD algorithm biologically related positions are selected in the absolute spectrum of a problem and a model is constructed with simple rules in order to characterise all protein classes. A detailed analysis for the subgroups extracted for each subtype of virus is shown below:

- For the H1N1 Influenza A subtype one rule with two features is obtained to describe this subtype, features 44 and 97. For feature 44, as Fig. 4(c) shows, the linguistic label *Low* is associated to points −0.0125, 0.0 and 0.0125, and for feature 97, as Fig. 4(e) shows, the linguistic label *Low* is associated to points −0.0785, 0.0006 and 0.0791. The SD results obtained for this virus are very good with all the examples covered with a 96.6% of success. In addition, the unusualness and significance values are very high which shows an unusual behaviour of these properties, enabling the experts to characterise this subtype of virus. The interpretability of this subgroup is excellent with one subgroup represented by only two variables.
- One subgroup with four features is obtained for the H2N2 Influenza A subtype to describe this virus with the features 9, 54, 153 and 217. For feature 9, as Fig. 4(b) shows, the linguistic label *Low* is associated to points −0.0333, 0.0015 and 0.0348, for feature 54, as Fig. 4(d) shows, the linguistic label *Low* is associated to points −0.0551, 0.0068 and 0.0619, for feature 153, as Fig. 4(h) shows, the linguistic label *Low* is associated to points −0.0424, 0.0002 and 0.0426, and finally, for feature 217, as Fig. 4(j) shows, the linguistic label *Low* is associated to −0.0334, 0.0 and 0.0334. All the examples for this subtype of virus are covered because the sensitivity is equal to 100.0%, with a 60.0% degree of success. The value of significance indicates a relative significance of this subtype of virus with respect to the others. Despite this subtype having a low number of instances the unusualness value is important.
- For the H3N2 Influenza A subtype two rules are obtained to describe this subtype. For the first rule feature 8 is used and for the second rule features 141, 207 and 219. For feature 8, as Fig. 4(a) shows, the linguistic label *Low* is associated to the points −0.0380, 0.006 and 0.0386, for feature 141, as Fig. 4(g) shows, the linguistic label *Low* is associated to −0.0243, 0.0 and 0.0243, for feature 207, as Fig. 4(i) shows, the linguistic label *Low* is associated to the points −0.0409, 0.0021 and 0.0430, and finally, for feature 219, as Fig. 4(k) shows, the linguistic label *Low* is associated to the points −0.04, 0.0 and 0.04. To represent this subtype of virus two different subgroups can be observed. On the one hand, a general subgroup with only one feature where all the examples for the subtypes are covered with 73.0% of proteins covered correctly and with excellent results in significance with respect to other subgroups. On the other hand, a more specific subgroup is obtained with three features where 99.5% of proteins are covered with 96.6% of success. This relationship between sensitivity and confidence yields a good rule for describing and classifying new instances of this type of virus.
- For the H5N1 Influenza A subtype one feature of the absolute spectra was created to classify this subtype, feature 115. For feature 115, as Fig. 4(f) shows, the linguistic label *Low* is associated to the points −0.0422, 0.0006 and 0.0428. This subgroup covers all the proteins of this subtype with a success rate of 67.7%. The results for significance and unusualness are interesting considering that this subtype has the lowest number of instances of the data set. The interpretability of this subgroup is excellent with one subgroup is extracted with only one variable.

7. Conclusions and future work

In this paper, the Influenza A virus problem is tackled through a SD algorithm which is able to provide novel knowledge to the experts. The main objective of the paper was to find interpretable knowledge in the Influenza A virus problem in order to describe unusual behaviour in several subtypes of this virus.

For this purpose, one of the most representative SD algorithms was applied, the NMEEF-SD algorithm. NMEEF-SD is based on an EFS which is suitable for extracting rules with few features, i.e. interpretable, in order to facilitate the comprehensibility of the subgroups because the algorithm also uses fuzzy logic to analyse the features. On the one hand the fuzzy logic with linguistic labels avoids a previous discretisation of the variables in the data set, and increases the interpretability of the knowledge extracted.

The results obtained for the algorithm show a good behaviour for this real-world problem from two points of view:

- Descriptive: The NMEEF-SD obtains representative subgroups for each subtype of the virus. These subgroups show an unusual and significant behaviour and also represent the total examples for each class, i.e. good values of sensitivity, with a good value of confidence.
- Predictive: the algorithm obtains a good level of precision in order to classify new proteins to be included in the data set. Furthermore, the rules extracted in order to classify new examples are very interpretable because the algorithm employs linguistic labels to represent the continuous features, and because the number of features for each subgroup is very low.

This paper offers the community a new point of view in the analysis of the Influenza A virus with a novel technique characterised by its interpretability, which obtains simple rules to represent different subtypes of the virus. In this way the model can classify an unknown protein sequence in a subtype of virus with only 11 features of the absolute spectrum instead of the whole spectrum, which consists of 256 features. Therefore the study shows similarities/differences between different subtypes of the Influenza A virus which can then help experts to the understanding of the Influenza A virus domain.

Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness under projects TIN2012-33856 FEDER founds, and by the Andalusian Research Plan under project TIC-3928 FEDER founds. This paper was written in the Centre for Computational Intelligence, De Montfort University Leicester, during the research visit of C.J. Carmona. Thanks to Prof. R. John and Prof. F. Chiclana for the invitation.

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, Journal of Molecular Biology 215 (3) (1990) 403–410.

[2] E. Pirogova, Q. Fang, E. Lazoura, I. Cosic, Analysis of amino acid parameters in the resonant recognition model, in: Proceedings of the 2nd International Conference on Bioelectromagnetism, 1998, pp. 71–72.

[3] I. Cosic, E. Pirogova, Bioactive peptide design using the Resonant Recognition Model, Nonlinear Biomedical Physics 1 (1) (2007) 7.

[4] I. Cosic, Macromolecular bioactivity: is it resonant interaction between macromolecules: theory and applications, IEEE Transactions on Bio-medical Engineering 41 (1994) 1101–1114.

[5] C. Chrysostomou, H. Seker, N. Aydin, P. Haris, Complex resonant recognition model in analysing Influenza A virus subtype protein sequences, in: Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine, 2010.

[6] W. Kloesgen, Explora: a multipattern and multistrategy discovery assistant, in: Advances in Knowledge Discovery and Data Mining, American Association for Artificial Intelligence, 1996, pp. 249–271.

[7] S. Wrobel, An algorithm for multi-relational discovery of subgroups, in: Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery, Vol. 1263 of LNAI, Springer, 1997, pp. 78–87.

[8] D. Gamberger, N. Lavrac, F. Zelezny, J. Tolar, Induction of comprehensible models for gene expression datasets by subgroup discovery methodology, Journal of Biomedical Informatics 37 (4) (2004) 269–284.

[9] N. Lavrac, Subgroup discovery techniques and applications, in: Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining, vol. 3518 of LNCS, Springer, 2005, pp. 2–14.

[10] I. Trajkovski, F. Zelezny, N. Lavrac, J. Tolar, Learning relational descriptions of differentially expressed gene groups, IEEE Transactions on Systems, Man, and Cybernetics, Part C 38 (1) (2008) 16–25.

[11] F. Herrera, C.J. Carmona, P. González, M.J. del Jesus, An overview on subgroup discovery: foundations and applications, Knowledge and Information Systems 29 (3) (2011) 495–525.

[12] F. Herrera, Genetic fuzzy systems: taxomony, current research trends and prospects, Evolutionary Intelligence 1 (2008) 27–46.

[13] L.A. Zadeh, The concept of a linguistic variable and its applications to approximate reasoning. Parts I, II, III, Information Science 8-9 (1975), 199-249,301-357,43-80.

[14] E. Hüllermeier, Fuzzy sets in machine learning and data mining, Applied Soft Computing 11 (2) (2011) 1493–1505.

[15] M.J. del Jesus, P. González, F. Herrera, M. Mesonero, Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing, IEEE Transactions on Fuzzy Systems 15 (4) (2007) 578–592.

[16] M.J. del Jesus, P. González, F. Herrera, Multiobjective Genetic Algorithm for Extracting Subgroup Discovery Fuzzy Rules, in: Proceedings of the IEEE Symposium on Computational Intelligence in Multicriteria Decision Making, IEEE Press, 2007, pp. 50–57.

[17] C.J. Carmona, P. González, M.J. del Jesus, F. Herrera, NMEEF-SD: Non-dominated multi-objective evolutionary algorithm for extracting fuzzy rules in subgroup discovery, IEEE Transactions on Fuzzy Systems 18 (5) (2010) 958–970.

[18] C.J. Carmona, P. González, M.J. del Jesus, M. Navío, L. Jiménez, Evolutionary fuzzy rule extraction for subgroup discovery in a psychiatric emergency department, Soft Computing 15 (12) (2011) 2435–2448.

[19] C.J. Carmona, P. González, M.J. del Jesus, S. Ventura, Subgroup discovery in an e-learning usage study based on Moodle, in: Proceedings of the International Conference of European Transnational Education, 2011, pp. 446–451.

[20] A. Moscona, Neuraminidase inhibitors for influenza, New England Journal of Medicine 353 (13) (2005) 1363.

[21] Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, D. Lipman, The influenza virus resource at the National Center for Biotechnology Information, Journal of Virology 82 (2) (2008) 596.

[22] D.M. Morens, J.K. Taubenberger, A.S. Fauci, The persistent legacy of the 1918 influenza virus, New England Journal of Medicine 361 (3) (2009) 225.

[23] M.M. Mukhtar, S.T. Rasool, D. Song, C. Zhu, Q. Hao, Y. Zhu, J. Wu, Origin of highly pathogenic H5N1 avian influenza virus in China and genetic characterisation of donor and recipient viruses, Journal of General Virology 88 (Part 11) (2007) 3094–3099.

[24] V. Veljkovic, N. Veljkovic, C.P. Muller, S. Mueller, S. Glisic, V. Perovic, H. Koehler, Characterization of conserved properties of hemagglutinin of H5N1 and human influenza viruses: possible consequences for therapy and infection control, BMC Structural Biology (2009) 9.

[25] V. Veljkovic, I. Cosic, B. Dimitrijevic, D. Lalovic, Is it possible to analyze DNA and protein sequences by the methods of digital signal processing? IEEE Transaction on Biomedical Engineering 32 (5) (1985) 337–341.

[26] K. Gopalakrishnan, R.H. Zadeh, K. Najarian, A. Darvish, Computational analysis and classification of p53 mutants according to primary structure, in: Proceedings of the IEEE Computational Systems Bioinformatics Conference, 2004, pp. 694–695.

[27] C. Chrysostomou, H. Seker, N. Aydin, Effects of windowing and zero-padding on complex resonant recognition model for protein sequence analysis, in: 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Boston (USA), 2011, pp. 4955–4958.

[28] R. Blackman, J.W. Tukey, The Measurement of Power Spectra: From the Point of View of Communications Engineering, Dover Publications, Mineola, NY, 1958.

[29] R.F. Henry, P.W.U. Graefe, Zero Padding as a Means of Improving Definition of Computed Spectra, Published for Environment Canada by Department of Energy, Mines and Resources, Marine Sciences Branch, 1971.

[30] D. Sundararaja, The Discrete Fourier Transform: Theory, Algorithms and Applications, World Scientific Pub Co Inc, 2001.

[31] S. Wrobel, Inductive Logic Programming for Knowledge Discovery in Databases, Springer-VerlagBerlin Heidelberg, New York, 2001, Ch. Relational Data Mining, pp. 74–101.

[32] M. Atzmueller, F. Puppe, H.P. Buscher, Towards knowledge-intensive subgroup discovery, in: Proceedings of the Lernen – Wissensentdeckung 1- Adaptivität 1- Fachgruppe Maschinelles Lernen, 2004, pp. 111–117.

[33] D. Gamberger, N. Lavrac, Active subgroup mining: a case study in coronary heart disease risk group detection, Artificial Intelligence in Medicine 28 (1) (2003) 27–57.

[34] W. Kloesgen, J. Zytkow, Handbook of Data Mining and Knowledge Discovery, Oxford University Press, Oxford, 2002.

[35] N. Lavrac, B. Cestnik, D. Gamberger, P.A. Flach, Decision support through subgroup discovery: three case studies and the lessons learned, Machine Learning 57 (1–2) (2004) 115–143.

[36] N. Lavrac, P.A. Flach, B. Zupan, Rule Evaluation Measures: A Unifying View, in: Proceedings of the 9th International Workshop on Inductive Logic Programming, vol. 1634 of LNCS, Springer, Bled, Slovenia, 1999, pp. 174–185.

[37] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. Verkamo, Fast discovery of association rules, in: U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in Knowledge Discovery and data mining, AAAI Press / MIT Press, Cambridge, CA, USA, 1996, pp. 307–328.

[38] A.E. Eiben, J.E. Smith, Introduction to Evolutionary Computation, Springer-Verlag Berlin Heidelberg, New York, 2003.

[39] T.M. Basu, N.K. Mahapatra, S.K. Mondal, A balanced solution of a fuzzy soft set based decision making problem in medical science, Applied Soft Computing 12 (10) (2012) 3260–3275.

[40] D. Chen, C. Gao, Soft computing methods applied to train station parking in urban rail transit, Applied Soft Computing 12 (2) (2012) 759–767.

[41] C.H. Tan, K.S. Yap, H.J. Yap, Application of geneticalgorithm for fuzzy rules optimization on semi expert judgment automation using Pittsburg approach, Applied Soft Computing 12 (8) (2012) 2166–2177.

[42] T. Bäck, D. Fogel, Z. Michalewicz, Handbook of Evolutionary Computation, Oxford University Press, Oxford, 1997.

[43] J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, ACM, NY, USA, 1975.

[44] T. Kumar-Paul, H. Iba, Prediction of cancer class with majority voting genetic programming classifier using gene expression data, IEEE/ACM Transactions on Computational Biology and Bioinformatics 6 (2) (2009) 353–367.

[45] K. Ghosein, S.F. Ghannadpour, A hybrid geneticalgorithm for multi-depot homogenous locomotive assignment with time windows, Applied Soft Computing 10 (1) (2010) 53–65.

[46] A. Singh, W.N. Bhukya, A hybrid geneticalgorithm for the minimum energy broadcast problem in wireless ad hoc networks, Applied Soft Computing 11 (1) (2011) 667–674.

[47] F. Fernández-Navarro, C. Hervás-Martínez, R. Ruiz, J.C. Riquelme, Evolutionary generalised radial basis function neural networks for improving prediction accuracy in gene classification using feature selection, Applied Soft Computing 12 (6) (2012) 1787–1800.

[48] K. Deb, A. Pratap, S. Agrawal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Transactions on Evolutionary Computation 6 (2) (2002) 182–197.