Subgroup Discovery through Evolutionary Fuzzy Systems applied to Bioinformatics problems

C. J. Carmona
Languages and Computer Technology Systems, Department of Civil Engineering, University of Burgos, 09006, Burgos (Spain)
D. Elizondo
School of Computer Science and Informatics, De Montfort University, The Gateway, Leicester, LE1 9BH, (United Kingdom)

Abstract

Subgroup discovery is a descriptive data mining technique using supervised learning. This paper presents a summary about the main properties and elements about subgroup discovery task. In addition, we will focus on the suitability and potential of the search performed by evolutionary algorithms in order to apply in the development of subgroup discovery algorithms, and in the use of fuzzy logic which is a soft computing technique very close to the human reasoning. The hybridisation of both techniques are well known as evolutionary fuzzy system.

The most relevant applications of evolutionary fuzzy systems for subgroup discovery in the bioinformatics domains are outlined in this work. Specifically, these algorithms are applied to a problem based on the Influenza A virus and the accute sore throat problem.

1. Introduction

Subgroup discovery (SD) is a descriptive data mining technique for describing unusual features with monitored properties of interest (Kloesgen, 1996, Wrobel, 1997). This task contributes interesting knowledge to the scientific community from two view-points, specifically both features those including the provision of interest and precision. SD has been included within the concept of Supervised Descriptive Rule Discovery (Kralj-Novak et al., 2009), together with further descriptive techniques such as emerging patterns (Dong and Li, 2005) and contrast set mining (Bay and Pazzani, 2001).

Differing SD algorithms have been implemented throughout the literature in order to solve SD tasks based on beam search such as CN2-SD (Lavrac et al., 2004a) or SD (Gamberger and Lavrac, 2002), exhaustive such as SD-Map (Atzmueller and Puppe, 2006), or genetic algorithms such as SDIGA (del Jesus et al., 2007b) and NMEEF-SD (Carmona et al., 2010a), amongst others.

This paper presents different applications of one type of SD algorithms to the bioinformatics domain. Specifically, the type of algorithms are evolutionary fuzzy systems (EFSs) (Herrera, 2008) based on evolutionary algorithms (Holland, 1975, Goldberg, 1989) and fuzzy logic (Zadeh, 1994). The applications are related to the Influenza A virus and the accute sore throat. The EFSs demonstrate a good behaviour in order to solve this type of problems.

The paper is organised as follows. Firstly, the Section 2 describes the main properties, elements and quality measures of the SD task. Section 3 presents the EFSs and the main algorithms based on EFSs presented for SD throughout the literature. Finally, Section 4 describes the main applications within the bioinformatics domain solved through EFSs for SD.

2. Subgroup discovery

In the following subsections the formal definition of the subgroup discovery task, the relation with other data mining tasks and the main elements of a SD algorithm are depicted.

2.1. Definition of subgroup discovery

The concept of SD was initially introduced by Kloesgen (Kloesgen, 1996) and Wrobel (Wrobel, 1997), and more formally defined by Siebes (Siebes, 1995) but using the name Data Surveying for the discovery of interesting subgroups. It can be defined as (Wrobel, 2001):

> *"In SD, 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 SD 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."*

SD attempts to search relations between different properties or variables of a set with respect to a target variable. Due to the fact that SD is focused in the extraction of relations with interesting characteristics, it is not necessary to obtain complete but partial relations. These relations are described in the form of individual rules.

Then, a rule (R), which consists of an induced subgroup description, can be formally defined as (Gamberger and Lavrac, 2002, Lavrac et al., 2004a):

$$R : Cond \ \rightarrow \ Target_{value}$$

where $Target_{value}$ is a value for the variable of interest (target variable) for the SD task (which also appears as $Class$ in the literature), and $Cond$ is commonly a conjunction of features (attribute-value pairs) which is able to describe an unusual statistical distribution with respect to the $Target_{value}$.

As an example, let D be a dataset with three variables $Age = \{Less\ than\ 25,\ 25\ to\ 60,\ More\ than\ 60\}$, $Sex = \{M,\ F\}$ and $Country = \{Spain,\ USA,\ France,\ German\}$, and a variable of interest target variable $Money = \{Poor,\ Normal,\ Rich\}$. Some possible rules containing subgroup descriptions are:

$$R_1 :\ (Age = Less\ than\ 25\ AND\ Country = German) \rightarrow\ Money = Rich$$
$$R_2 :\ (Age = More\ than\ 60\ AND\ Sex = F) \rightarrow\ Money = Normal$$

where rule R_1 represents a subgroup of German people with less than 25 years old for which the probability of being rich is unusually high with respect to the rest of the population, and rule R_2 represents that women with more than 60 years old are more likely to have a normal economy than the rest of the population.

SD is somewhere halfway between predictive and descriptive induction, and its goal is to generate in a single and interpretable way subgroups to describe relations between independent variables and a certain value of the target variable. The algorithms for this task must generate subgroups for each value of the target variable. Therefore, an execution for each value of the variable must be performed.

A rule for SD is represented in Fig. 1, where two values for the target variable can be found ($Target_{value} = x$ and $Target_{value} = o$). In this representation a subgroup for the first value of the target variable can be observed, where the rule attempts to cover a high number of objects with a single function: a circle. As can be observed the subgroup does not cover all the examples for the target value x even the examples covered are not positive in all the cases, but the form of this function is uniform and very interpretable with respect others. In this way the algorithm achieves a reduction of the complexity. Furthermore, the true positive rate for the value of the target variable is high, with a value of 75%.

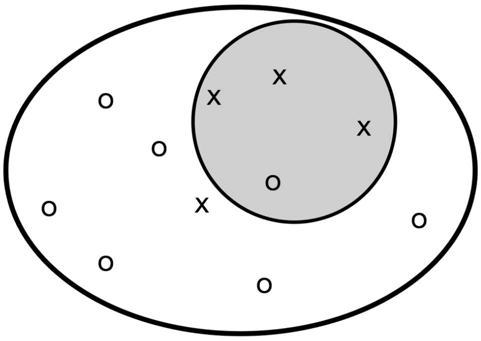


Figure 1: Representation of a subgroup discovery rule with respect to a value (x) of the target variable

The SD task 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 complex and precise.

Different elements can be considered the most important when a SD approach must be applied. These elements are defined below (Atzmueller et al., 2004):

- *Type of the target variable*. Different types for the variable can be found: binary, nominal or numeric. For each one different analyses can be applied considering the target variable as a dimension of the reality to study.

- *Description language*. The representation of the subgroups must be suitable for obtaining interesting rules. These rules must be simple and therefore are represented as attribute-value pairs in conjunctive or disjunctive normal form in general. Furthermore, the values of the variables can be represented as positive and/or negative, through fuzzy logic, or through the use of inequality or equality and so on.

- *Quality measures*. These are a key factor for the extraction of knowledge because the interest obtained depends directly on them. Furthermore, quality measures provide the expert with the importance and interest of the subgroups obtained. Different quality measures have been presented in the specialised bibliography (Gamberger and Lavrac, 2003, Kloesgen, 1996, Kloesgen and Zytkow, 2002, Lavrac et al., 2004a), but there is no consensus about which are the most suitable for use in SD.

- *Search strategy*. This is very important, since the dimension of the search space has an exponential relation to the number of features and values considered. Different strategies have been used up to the moment, for example beam search, evolutionary algorithms, search in multi-relational spaces, etc.

2.2. Quality measures

A wide number of quality measures have been presented in the SD literature both to guide the search process in order to find the best SD rules and to measure the quality of the SD rule set finally obtained (Kloesgen, 1996, Lavrac et al., 2004b). The most common quality measures used in SD can be classified by their main objective such as:

- Complexity measures, related to the interpretability of the subgroups, i.e. to the simplicity of the knowledge extracted.
- Generality measures, used to quantify the quality of individual rules according to the individual patterns of interest covered.
- Precision measures, showing the precision of the subgroups.
- Interest measures, intended for selecting and ranking patterns according to their potential interest to the user.
- Hybrid, that attempt to obtain a good trade-off between different objectives.

Table 1 summarises the *Quality measures* most used in SD (Herrera et al., 2011) and their main characteristics.

Table 1: Classification of the quality measures used in subgroup discovery

Quality measure	*C*	*G*	*P*	*I*
Number of rules	X			
Number of variables	X			
Coverage (Lavrac et al., 2004b)		X		
Support (Lavrac et al., 2004b)		X		
Confidence (Agrawal et al., 1996)			X	
Precision measure Q_c (Gamberger and Lavrac, 2002)			X	
Precision measure Q_g (Kloesgen, 1996)			X	
Q_g-Weight (Gamberger and Lavrac, 2002)			X	
Interest (Noda et al., 1999)				X
Novelty (Wrobel, 1997)				X
Lift (Brin et al., 1997)				X
Significance (Kloesgen, 1996)				X
Sensitivity (Kloesgen, 1996)		X	X	
False Alarm (Gamberger and Lavrac, 2002)		X	X	
Specificity (Kloesgen, 1996)		X	X	
Unusualness (Lavrac et al., 1999)		X	X	X
Piatetstky-Shapiro (Grosskreutz et al., 2008)		X	X	X

C=Complexity, G=Generality, P=Precision and I=Interest

According to the SD concept the obtaining of interesting, simple and interpretable subgroups, covering the majority of the examples of the interest property (target variable) is desirable. Considering this definition and the analysis of the different quality measures used in the literature, we propose three guidelines in order to establish the type of measure more suitable, to guide the search process and to analyse the quality of the subgroups obtained by any SD algorithm:

- *Interpretability*. A SD proposal must obtain few rules containing a low number of variables in the antecedent part in order to help to the experts to understand and use the extracted knowledge, i.e. simple and interpretable subgroups are preferred in SD task.
- *Relation sensitivity-confidence*. A SD algorithm must obtain results with a good precision, where the majority examples covered belong to the target variable, i.e. it must achieve the best possible relation between sensitivity and confidence. Both quality measures are primordial in order to provide subgroups to the experts that cover the higher number of described correctly examples. It is difficult for the algorithms to obtain this compromise due to the loss suffered by a measure when trying to increase the other.
- *Novelty*. A SD model must contribute novel knowledge, providing the experts with information in order to describe unusual and interesting behaviour within the data. This objective could be measured with a wide number of quality measures as novelty, interest or significance, among others. Nevertheless, it is important to highlight the utility of the unusualness to measure this objective because it contributes with generality and confidence to the problem. Moreover, this quality measure is widely used in the specialised bibliography.

It can be considered that the main purpose of a SD algorithm is to find a good trade-off between these three guidelines because this lead to the obtaining of good results in a wide number of quality measures and not only in the ones used in the search process.

3. Evolutionary fuzzy systems

Computational Intelligence techniques such as artificial neural networks (Rojas, 1996), fuzzy logic (Yager and Filev, 1994), and genetic algorithms (Holland, 1975, Goldberg, 1989) are popular research subjects, since they can deal with complex engineering problems which are difficult to solve by classical methods (Konar, 2005).

Hybrid approaches have attracted considerable attention in the Computational Intelligence community. One of the most popular approaches is the hybridization between fuzzy logic and GAs leading to genetic fuzzy systems (Cordón et al., 2001). A GFS is basically a fuzzy system augmented by a learning process based on evolutionary computation, which includes genetic algorithms, genetic programming, and evolutionary strategies, among other evolutionary algorithms (EAs) (Eiben and Smith, 2003). This concepts is extended to the EFSs (Herrera, 2008).

Fuzzy systems are one of the most important areas for the application of the Fuzzy Set Theory (Zadeh, 1965, 1975). Usually it is considered a model structure in the form of fuzzy rule based systems (FRBSs). FRBSs constitute an extension to classical rule-based systems, because they deal with ”IF-THEN” rules, whose antecedents and consequents are composed of fuzzy logic statements, instead of classical ones. They have demonstrated their ability for control problems (Palm et al., 1997), modelling (Pedrycz, 1996), classification or data mining (Kuncheva, 2000) in a huge number of applications.

The automatic definition of an FRBS can be seen as an optimization or search problem, and EAs are a well known and widely used global search technique with the ability to explore a large search space for suitable solutions only requiring a performance measure. In addition to their ability to find near optimal solutions in complex search spaces, the generic code structure and independent performance features of EAs make them suitable candidates to incorporate a priori knowledge. In the case of FRBSs, this a priori knowledge may be in the form of linguistic variables, fuzzy membership function parameters, fuzzy rules, number of rules, etc. These capabilities extended the use of GAs in the development of a wide range of approaches for designing FRBSs over the last few years.

The SD is focused on the genetic rule learning where most of the approaches proposed to automatically learn the knowledge base from numerical information have focused on the rule base learning, using a predefined data base. The usual way to define this DB involves choosing a number of linguistic terms for each linguistic variable (an odd number between 3 and 9, which is usually the same for all the variables) and setting the values of the system parameters by an uniform distribution of the linguistic terms into the variable universe of discourse.

Following subsections present the EFSs for SD presented throughout the literature, as far as we know.

3.1. SDIGA

SDIGA(del Jesus et al., 2007b) is an evolutionary fuzzy system (Herrera, 2008) because it uses a knowledge representation fuzzy rules and evolutionary computation as a learning process. It is interesting to remark that SDIGA searches for rules for each value of the target variable, i.e. the consequent is not represented in the chromosome but is fixed.

This algorithm follows the IRL approach where the solution of each iteration is the best individual obtained and the global solution is formed by the best individuals obtained in the different runs. The representation of the individuals is performed through the "*Chromosome = Rule*" approach and the core of SDIGA is an EA using a post-processing step based on a local search. This hybrid algorithm extracts one simple and interpretable fuzzy rule with an adequate level of support and confidence. The algorithm model can use fuzzy canonical or DNF rules with a predefined set of linguistic labels.

This algorithm is included in an iterative process for the extraction of different rules. In this way, algorithm marks examples cover for rules to prevent a new rule being obtained which covers exactly the same examples in the following runs. Algorithm is obtaining rules while the generated rules reach a minimum level of confidence and give information on areas of the search space in which there are examples not described by the rules generated in previous iterations. The rule is improved in a post-processing phase throughout a hill-climbing process, which modifies the rule in order to increase the degree of support.

The fitness is an aggregation function where the selection of the quality measures like coverage, significance, unusualness, accuracy, sensitivity, crisp support, fuzzy support, crisp confidence and fuzzy confidence is determined by the user. The number of objectives within the weighted aggregation function are between 1 and 3.

SDIGA is implemented in the KEEL software tool (Alcalá-Fdez et al., 2011, Alcalá-Fdez et al., 2009) and its operation scheme can be observed in Fig. 2. This algorithm has been applied in order to search for unusual relationships in different real-world problems such as:

- Marketing, which was analysed in (del Jesus et al., 2007b). The main objective of this paper was to extract conclusions from the information on previous trade fairs to determine the relationship between the trade fair planning variables and the success of the stand. SDIGA was applied in order to extract information of interest about each of the three efficiency groups of stands: low, medium and high efficiency.

- Medicine, for the discovery and description of patients patterns in a psychiatric emergency department (Carmona et al., 2011a). In this work were presented rules describing relationships between the different variables stored for each patient and the arrival time divided in different periods: day, afternoon and night.

- E-learning, which was analysed in different works (Carmona et al., 2010b, 2011b). The main objective was to determine behaviour patterns for the students in e-learning platforms at the University of Cordoba. Both papers analyse possible relations between the usage of complementary activities of a course and the final marks obtained by the students. The final mark is used as the variable to characterize, using the different marks to divide the data into classes and codifying them as values of the consequent of the rules.

```
BEGIN
Set of rules is empty
repeat
  repeat
    Generate P(0)
    Evaluate P(0)
    repeat
      Include the best individual in P(nGen+1)
      Complete P(nGen+1): Crossover and Mutation of individuals from P(nGen)
      Evaluate P(nGen+1)
    until Number of evaluations is not reached
    Obtain the best rule R
    Local search R
    if Confidence(R) $\geq$ Minimum Confidence and R represents new examples then
      Set of rules $\bigcup$ R
      Mark the set of examples covered by R
    end if
    Set of rules $\bigcup$ R
  until Confidence(R) $\leq$ Minimum Confidence and R not represents new examples
until Not $Target_{value}$
END
```

Figure 2: Operation scheme of SDIGA algorithm

3.2. MESDIF

MESDIF (del Jesus et al., 2007a) is a multiobjective EA is an evolutionary fuzzy system based on the SPEA2 approach (Zitzler et al., 2002). It applies the concepts of elitism in the rule selection (using a secondary or elite population) and the search for optimal solutions in the Pareto front. In order to preserve the diversity at a phenotypic level the algorithm uses a niches technique which considers the proximity in values of the objectives and an additional objective based on novelty to promote rules which give information on examples not described by other rules of the population.

The rule induction process obtains rules with high predictive accuracy and which are comprehensible and interesting. In this proposal, the user can choose between a wide number of quality measures (coverage, significance, unusualness, accuracy, sensitivity, support and confidence) to maximise all the defined objectives.

One of the most important aspects of MESDIF is the obtention of results for all the values of the target variable. It returns the individuals of the elite population for each value, whose size is defined by the user.

The algorithm uses the "*Chromosome = Rule*" approach. The multiobjective EA discovers fuzzy rules whose consequent is prefixed to one of the possible values of the target feature. Therefore, all the individuals of the population are associated with the same value of the target variable, and so the chromosome only represents the antecedent of the rule.

MESDIF is implemented in KEEL (Alcalá-Fdez et al., 2011, Alcalá-Fdez et al., 2009) and its operation scheme can be observed in Fig. 3. This algorithm was applied in real-world problems such as marketing (Berlanga et al., 2006), medicine (Carmona et al., 2011a) and e-learning (Carmona et al., 2010b, 2011b).

```
BEGIN
Set of rules is empty
repeat
  Initialise counters
  Generate an initial population P(0) and create an empty elite population P'(0).
  repeat
    Calculate Fitness for P(nGen)
    Copy non-dominated individuals in P'(nGen)
    Generate P(nGen+1): Select, Crossover and Mutation from P(nGen)
  until Number of evaluations is reached
  Set of rules $\bigcup$ Non-dominated individuals P(nGen)
until Not $Target_{value}$
```

Figure 3: Operation scheme of MESDIF algorithm

3.3. NMEEFSD

NMEEF-SD (Carmona et al., 2010a) is a multiobjective evolutionary fuzzy system based on NSGA-II (Deb et al., 2002). NMEEF-SD codifies each candidate solution according to 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. Therefore, the algorithm must be executed as many times as the number of different values the target variable contains. With respect to the representation of the rules NMEEF-SD can use canonical or DNF rules.

As the general objective of NMEEF-SD is to obtain a set of rules, which should be general and accurate, the algorithm includes components which enhance these characteristics. In particular, diversity is enhanced in the population using a new operator to perform a re-initialisation based on coverage, in addition to a niching technique (the crowding distance in the selection operator). On the other hand, in order to promote generalisation, as well as the objectives considered in the evolutionary approach, the algorithm includes operators of biased initialisation and biased mutation. Finally, to ensure accuracy, in addition to the objectives NMEEF-SD returns as its final solution those rules which reach a predetermined confidence threshold.

NMEEF-SD allows to choose between two and three quality measures as objectives of the evolutionary process in order to obtain relevant subgroups, between: coverage, significance, unusualness, accuracy, sensitivity, support and confidence. It is also implemented in KEEL (Alcalá-Fdez et al., 2011, Alcalá-Fdez et al., 2009) and the operation scheme can be observed in Fig. 4. Recent applications to real-problems with an EA for SD have been analysed through this algorithm in different problems such as:

- E-learning (Carmona et al., 2011b), where a description of possible relationships between the use of the e-learning platform and marks obtained by the students were analysed. NMEEF-SD obtained a comprehensive set of subgroups employing a low number of variables with the highest unusualness. Subgroups obtained allowed to the teachers to take decisions about course activities to improve the performance of their students.

- E-commerce (Carmona et al., 2012), where the main objective of this paper was to analyse the usage of customers in a website based on the sell of olive oil in order to improve its design. Conclusions obtained have helped to the webmaster team to improve the design of the website.

- Bioinformatic (Carmona et al., 2013a), which presents an interesting study in this domain in order to find interpretable knowledge in the Influenza A virus problem and describe unusual behaviour in several subtypes of this virus. The results of NMEEF-SD offer the community a new point of view in the analysis of the Influenza A virus by its interpretability, which obtains simple rules to represent different subtypes of the virus.

- Concentrating photovoltaic technology was analysed in (Carmona et al., 2013b). This technology is an alternative to the conventional photovoltaic for electric generation. It produces electricity in a cheaper way by means of high efficiency multi-junction solar cells. The main objective was to describe the main external variables which improve the performance of the solar cells. The results confirmed some relationships between atmospheric variables and maximum power as well as new knowledge.

3.4. FuGePSD

FuGePSD (Carmona et al., 2015) commences from an initial population generated in a random manner where individuals are represented through the “chromosome=individual” approach including both the antecedent and the consequent of the rule. Specifically, FuGePSD employs the genetic cooperative-competition approach where rules of the population cooperate and compete between them in order to obtain the optimal solution. The inclusion of the target variable in the representation is often an advantage with respect to alternative EAs available for SD, since while FuGePSD is executed only once obtaining rules for the different values of the target variable. However, the remaining proposals based on EAs are required to be executed once for each value of the target variable. It is also important to remark that individuals have variable length just like population with a variable number of individuals throughout the evolutionary process. In this way, FuGePSD is able to obtain rules with different number of variables in the antecedent part of

```
BEGIN
Set of rules is empty
repeat
   Generate P(0)
   repeat
      Generate offspring Q(nGen) through operators in P(nGen)
      Join P(nGen) and Q(nGen) in R(nGen)
      Generate all non-dominated fronts from R(nGen)
      if the Pareto front evolves then
         Complete P(nGen+1) with fronts in order
      else
         Apply Re-initialisation based on coverage in P(nGen+1)
      end if
   until Number of evaluations is reached
   Set of rules $\bigcup$ Pareto front P(nGen)
until Not $Target_{value}$
END
```

Figure 4: Operation scheme of NMEEF-SD algorithm

the rule associated to the complexity of the subgroup to describe. On the other hand, the initial number of rules generated for the problem is adapted throughout the evolutionary process with respect to the problem to solve through different operators.

FuGePSD evolves with the generation of offspring populations through the application of several genetic operators. This population is generated with the same size than the parent population with respect to the number of individuals. Both populations are joined in a new population, in which the token competition operator is applied. This operator is crucial to the functioning of the algorithm in order to obtain diverse subgroups.

As can be observed in pseudo code of the Algorithm 5, this evolutionary process is controlled through the number of generations. It is important to note that for each generation, both individuals and population are evaluated in a separate manner, since in view of the use of the cooperative-competitive approach it is necessary to evaluate the individuals and populations through two independent fitness functions. Hence, individuals compete between themselves with respect to a local fitness, and cooperate in order to obtain a population which is more adapted to the problem. Once the evolutionary process has finished, the algorithm performs a screening function to obtain rules only with values greater than a threshold of sensitivity and confidence. These thresholds can be modified through external parameters providing to the experts an algorithm more adaptable to complex problems. In general, these thresholds should be configured upper than 60% level because subgroups obtained must be precise and general and both quality measures are ideal to meet these objectives. With these values, we may secure the extraction of interesting and effective rules for the SD task presented. The screening function is applied in the best population (*BestPop*) obtained throughout the complete evolutionary process.

4. Applications in bioinformatics domain

This section presents the use of EFSs in order to analyse bioinformatics problems through SD algorithms such as NMEEFSD and FuGePSD. Specifically, in Section 4.1 the analysis for

```
BEGIN
Set of rules is empty
Generate P(0)
Evaluate P(0)
$BestPop \longleftarrow$ P(0)
repeat
  Generate Q(nGen) through operators in P(nGen)
  Evaluate Q(nGen)
  Join P(nGen) and Q(nGen) in JoinPop(nGen)
  P(nGen) $\longleftarrow$ Token Competition from JoinPop(nGen)
  if $P(0).Fitness > BestPop.Fitness$ then
    BestPop $\longleftarrow$ P(nGen)
  end if
until Number of generations is reached
Set of rules $\bigcup$ Screening Function from BestPop
END
```

Figure 5: Operation scheme of FuGePSD algorithm

the Influenza A virus (Carmona et al., 2013a) is presented and Section 4.2 describes an acute sore throat problem (Carmona et al., 2015) analysed through the FuGePSD algorithm.

4.1. Influenza A virus

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, 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 (Moscona, 2005). 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 (Bao et al., 2008). 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 (Morens, D.M. and Taubenberger, J.K. and Fauci, A.S., 2009). 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 (Morens, D.M. and Taubenberger, J.K. and Fauci, A.S., 2009). 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 (Morens, D.M. and Taubenberger, J.K. and Fauci, A.S., 2009). 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' (Mukhtar et al., 2007), where H5N1 retains the NA gene from the avian H1N1 virus.

Percentage identity is a measurement used to determine the similarity between protein sequences. By using CLUSTALW, a freely available online tool (Morens, D.M. and Taubenberger, J.K. and Fauci, A.S., 2009), pairwise percent identity of all the subtype Influenza NA genes was calculated. Table 2 shows the average percentage identity between all the classes.

Table 2: Average Percent Identity

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

As Table 2 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.

4.1.1. 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 (Pirogova et al., 1998, Cosic and Pirogova, 2007, Cosic, 1994) and Complex Resonant Recognition Model (Chrysostomou et al., 2010). Previous studies (Veljkovic et al., 2009) 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) (Veljkovic et al., 1985, Gopalakrishnan et al., 2004) amino acid index as shown in Table 3 is used to express protein sequences in order to numerical sequences to be able to apply DFT.

4.1.2. 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 (Chrysostomou et al., 2011). Before applying DFT to the protein sequences, zero-padding and windowing, used in signal processing needs to be considered.

Table 3: 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

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 (Blackman and Tukey, 1958) is selected, and can be computed using Eq. 1.

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

The second technique under investigation is the zero-padding (Henry and Graefe, 1971, Sundararaja., 2001) 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.

4.1.3. Discrete Fourier Transform

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

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

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:

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

Table 4: 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

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:

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

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 equation 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.

4.1.4. Experimental study

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 4.

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 ($\sharp Rules$), number of variables ($\sharp Vars$), significance ($SIGN$), unusualness ($UNUS$), sensitivity ($SENS$) and confidence ($CONF$).

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 5. The best results for each quality measure are highlighted.

Table 5: Results obtained for the NMEEF-SD algorithm in the experimental study for the Influenza A virus problem

LLs	Min_{Cnf}	$\sharp Rules$	$\sharp Vars$	$SIGN$	$UNUS$	$SENS$	$CONF$
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

In Table 5 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 6, 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.

Table 6: 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}	$Class$	$\sharp Rules$	$\sharp Vars$	$SIGN$	$UNUS$	$SENS$	$CONF$
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

As mentioned previously in the analysis of Table 5 and with the results shown in Table 6, 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

Table 7: 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

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 $Class\ H1N1$ 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%.

- 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 7 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 8 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.

Table 8: 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 9: 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

The total accuracy for the complete data set is of 0.872 ± 0.051 for Table 7, and 0.797 ± 0.069 for Table 8. 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 2. 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.

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 9 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 9 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.

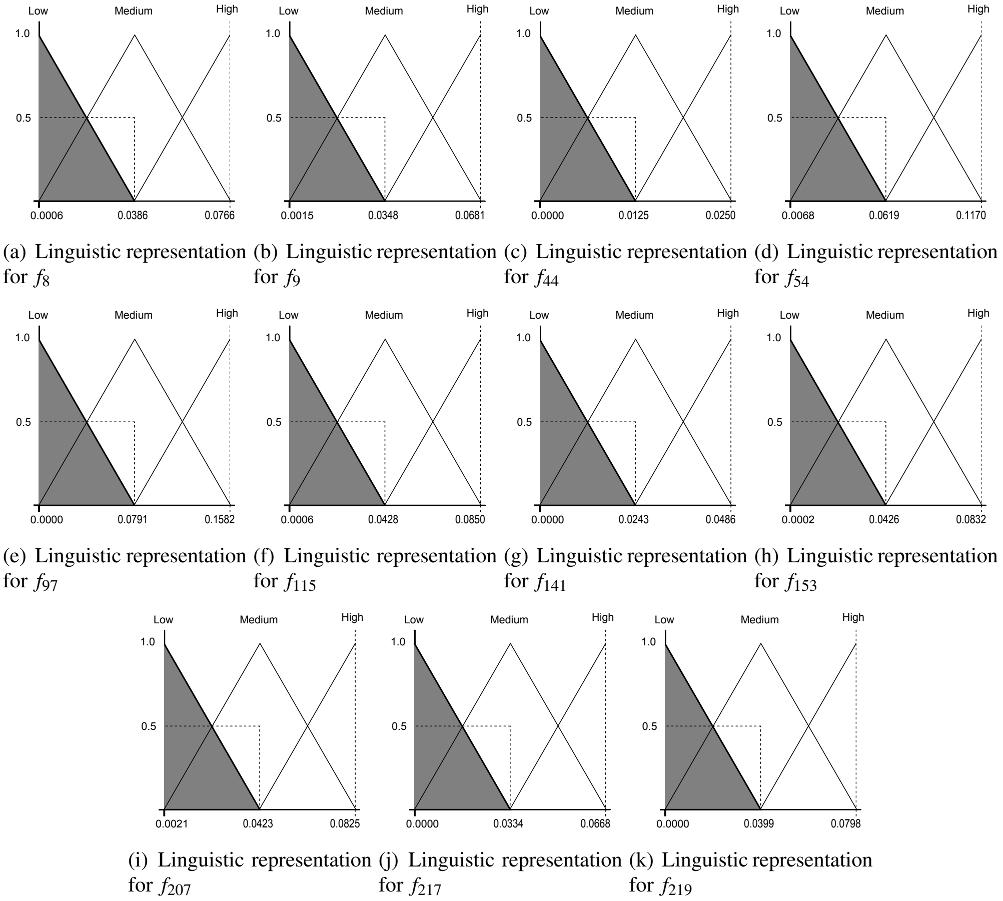
Other methods that uses signal processing techniques to extract biologically related features in order to characterise protein sequences like the Resonant Recognition Model in the HA gene (Veljkovic et al., 2009) and the Complex Resonant Recognition for the NA gene (Chrysostomou et al., 2010) use informational 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 peak (Chrysostomou et al., 2010). By using the NMEEF-SD algorithm simple rules, as table 9 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 figure 6(c) shows, the linguistic label *Low* is associated to points -0.0125, 0.0 and 0.0125, and for feature 97, as figure 6(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 figure 6(b) shows, the linguistic label *Low* is associated to points -0.0333, 0.0015 and 0.0348, for feature 54, as figure 6(d) shows, the linguistic label *Low* is associated to points -0.0551, 0.0068 and 0.0619, for feature 153, as figure 6(h) shows, the linguistic label *Low* is associated to points -0.0424, 0.0002 and 0.0426, and finally, for feature 217, as figure 6(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 figure 6(a) shows, the linguistic label *Low* is associated to the points -0.0380, 0.006 and 0.0386, for feature 141, as figure 6(g) shows, the linguistic label *Low* is associated to -0.0243, 0.0 and 0.0243, for feature 207, as figure 6(i) shows, the linguistic label *Low* is associated to the points -0.0409, 0.0021 and 0.0430, and finally, for feature 219, as figure 6(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 figure 6(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.



(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 for f_{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}

Figure 6: Linguistic representations of the continuous feature of the model extracted by the NMEEF-SD algorithm

4.2. Acute sore throat

Sore throat (sometimes known as 'pharyngitis' or 'tonsilitis') is an acute upper respiratory tract infection that impinges on the throat's respiratory mucosa, and can be linked with fever, headache and general malaise. Moreover, acute otitis media, acute sinusitis and peritonsillar abscess represent suppurative complications of this condition, predominantly the first of these. 85-95% of adult acute sore throat conditions are ascribable to viruses, as are 70% of those in children aged 5-16 years (and 95% of those in children aged < 5 years) (Worrall, 2006). However, the remainder arise from a bacterial source [predominantly group A β-hemolytic streptococcus (GABHS)]; clinically, the four most valuable features to identify in the diagnosis of sore throat diseases which are caused by GABHS are enlarged submandibular glands, the presence of a throat exudate and rhinorrhea (runny nose), together with the absence of fever and cough (Bisno et al., 2002, Worrall, 2006).

In this case study, we have coupled FuGePSD with high field proton (^{1}H) NMR spectroscopy analysis in order to recognise salivary biomolecule signatures which are characteristic of viral -(and, if applicable, bacterial)- induced acute sore throat conditions in humans. Specifically, healthy and clinically-diagnosed patients with acute sore throat conditions patients are analysed through FuGePSD in a problem with more than 200 variables and 500 instances. The main goal is to describe and characterise (from the point of view of SD) this problem with respect to the condition of the patients: healthy (control) and sore throat, i.e. with respect to two values for the target variable.

The applications of high field proton ^{1}H NMR spectroscopy to the detection and quantification of biomolecules present in complex biological fluids offers many advantages over other alternatives as can be observed in (Claxson et al., 1999, Grootveld et al., 1998, 1996, Silwood et al., 2002).

This case study has been performed in different stages: Firstly, patients were selected and data were collected as described in Section 4.2.1. Next, a preprocessing stage was applied to human saliva samples in order to obtain a data matrix for applying FuGePSD proposal in Section 4.2.2. Finally, Section 4.2.3 presents the analysis and results obtained with the FuGePSD algorithm.

4.2.1. Collection of human saliva samples

A series of patients with a clinically-diagnosed acute sore throat condition (n = 50) and healthy, non-medically-compromised age-matched controls (n = 50) were recruited to the study, the latter serving as essential controls. All of them were required to fully complete a participant questionnaire with both personal and medical information such as age, gender, body mass index, cough, rhinitis, fever history, etc. All participants were also instructed not to receive any form of medication during the 5-day trial. This investigation was performed by Professor Grootveld's research group, and full ethical approval for it was granted by the University of Bolton's Research Ethics Committee.

For the ^{1}H NMR data acquired we primarily implemented a rigorous analysis-of-variance (ANOVA)-based experimental design. This procedure was principally aimed at determining the significance of the 'Between-Disease Group' component of variance (and further ones involved) for the intensities of ^{1}H NMR intelligently-selected bucket signals which remained in the spectrum following the spectral editing process described below. A bucket is considered as an input variable in the dataset to analyse through the FuGePSD algorithm.

The experimental design selected was a combination of completely randomised with a randomised block design: mixed model with the 'Between-Participants' component of variance

Table 10: Experimental design for the analysis of each dataset of ^{1}H NMR ISB integration intensities, representing a combination of completely randomised with a randomised block design: mixed model with Participants (n=50 per Group) 'nested' within each of the two Disease Classification Groups

Source of variation	Levels	Degrees of freedom	Nature	Parameters Estimated
Between disease classifications	2	1	Fixed	$\sigma^2 + 5\sigma_{P(D)^2} + 250K_{D^2}$
Between participants	100	98	Random	$\sigma^2 + 5\sigma_{P(D)^2}$
Sampling days-withing-participants	5 per Volunteer	4	Sequentially-Fixed	$\sigma^2 + 100K_{S^2}$
Error (Residual)	n/a	396	n/a	σ^2
Total	n/a	499	n/a	n/a

(n=50 per Group) 'nested' within Disease Classification Group (Table 10). This model was preliminarily employed to probe the prognostic/diagnostic specificity of each 'Intelligently-Selected' Chemical Shift Bucket (ISB). Hence, this design allowed the study of each of these sources of variation simultaneously. For this ANOVA model, the complete dataset was log_{10}-transformed prior to analysis in order to satisfy assumptions of normality, variance homogeneity and additivity, etc. Saliva specimens were collected from each participant immediately after awakening in the morning as previously described in (Silwood et al., 2002). These 5 samplings took place throughout an intensive one-week period (Monday-Friday), and they were instructed to collect all saliva available in order to avoid interferences.

4.2.2. ^{1}H NMR analysis of human salivary supernatants

The preparation of human saliva samples for ^{1}H NMR analysis was performed as previously described (Silwood et al., 2002). Single-pulse ^{1}H NMR spectra of human salivary supernatant specimens were acquired on a Bruker Avance AM-600 spectrometer operating at a frequency of 600.13 MHz as described previously (Lemanska et al., 2011, Silwood et al., 2002, Wongravee et al., 2010), as were both one- and two-dimensional ^{1}H-^{1}H COSY and TOCSY spectra.

Main objective in these stages was to obtain a ^{1}H NMR data matrix with a spectra for each saliva specimen with different input variables. In this way, a matrix with 500 spectra and 209 intelligent chemical shift buckets (ISB input variables) generated via the application of macro procedures for line-broadening, zero-filling, Fourier-transformation and phase and baseline corrections, followed by the application of a separate macro for the 'Intelligent Bucketing' processing sub-routine is obtained; all procedures were performed with the ACD/Labs 1D NMR Manager software package (ACD/Labs, Toronto, Canada M5C 1T4). These buckets are selected through the employment of an algorithm designed to make critical divisional decisions, i.e. those which define precisely the loci of bucket divisions with regard to an optimised selection of 'resonance-specific' ones (B. Lefebvre, Intelligent Bucketing for Metabonomics, ACD/Labs Technical Note, 2004). This strategy generated one global table of 'intelligently-selected bucket' (ISB) intensities. Chemical shift buckets containing less than 1% of the maximum summed intensity were removed from the dataset (since they may contain spectral 'noise').

After removal of the intense H_2O resonance (δ=4.50-5.10 ppm), together with those arising from ethanol [centred at δ=1.21 (t) and 3.66 ppm (q)], all ISB variables, or, where indicated were incorporated into the dataset for analysis with FuGePSD. All chemical shift bucket intensity values were normalised to that of the pre-added TSP internal standard (of fixed concentration).

Table 11: Subgroups obtained in the case study through the FuGePSD method

Sb		UNUS	SENS	FCNF
1	IF ISB (6.31-6.33) = Low AND ISB (0.60-0.62) = Medium AND ISB (1.36-1.40) = Medium AND ISB (3.55-3.61) = Medium AND ISB (6.83-6.88) = Medium THEN Control	0.0167	0.8917	0.6854
2	IF ISB (2.22-2.27) = Medium AND ISB (2.29-2.31) = Medium AND ISB (2.78-2.83) = Medium AND ISB (5.66-5.69) = Medium THEN Sore Throat	0.0260	0.6958	0.9772
3	IF ISB (2.22-2.27) = Medium AND ISB (2.78-2.83) = Medium AND ISB (5.66-5.69) = Medium THEN Sore Throat	0.0271	0.7000	0.9259
4	IF ISB (2.22-2.27) = Medium AND ISB (5.66-5.69) = Medium AND ISB (8.37-8.42) = Medium THEN Sore Throat	0.0448	0.6625	0.8859
5	IF ISB (2.22-2.27) = Medium AND ISB (5.66-5.69) = Medium THEN Sore Throat	0.0290	0.7292	0.8154

4.2.3. Extraction of subgroup discovery by FuGePSD

Finally, the application of FuGePSD is performed on a dataset with 500 instances and 209 variables, and it is very important to note that buckets or input variables have a real domain. In this way, the use of this algorithm is relevant within SD task because as we have presented in the previous section, FuGePSD obtains the best results for these types of problems through the correct use of fuzzy logic.

Application of the FuGePSD method to the analysis of the salivary ^{1}H NMR dataset is performed with the standard parameters considered by authors but using as local fitness the fuzzy confidence because the main goal of the experts is to obtain accurate subgroups and using three linguistic labels.

Results acquired served to segregate the total number of saliva specimens into five subgroups: four to describe active sore throat saliva specimens and one for saliva specimens corresponding to healthy patients can be observed in Table 11, where the representation of the subgroups and their values for the quality measures are presented. FuGePSD is able to obtain subgroups with a low number of variables to describe both values for the target variable highlighting the values in unusualness and trade-off sensitivity-confidence. Subgroups are precise in general with an average confidence equal to 85.78%. In addition, with these subgroups the support reached for the algorithm is very close to the total of examples (95%). The metabolic assignment for each ISB is shown in the foot-table.

Furthermore, for each subgroup an statistical analysis for each ISB is performed shown the *Metabolic Assignment*, the *Sign* of classification mean difference between both target values, and the *ANOVA pValue* in Table 12. Valuable biomarker features identified were proteins, including those with relatively intense tyrosine residue resonances, acetoin and glycine, whereas those for the four sore throat disease classifications included 5-aminovalerate and the amino acid L-aspartate. The identity of the 5-aminovalerate signals (i.e., those coupled to the intense δ=2.24 ppm one) were confirmed via the acquisition of both 1D and 2D COSY ^{1}H/^{1}H-^{1}H NMR profiles

Table 12: Sore throat disease subgroups detectable via application of the FuGePSD method

Sb	ISB (ppm)	^{1}H NMR Resonance Mult.	Metabolic Assignment	Sign of Class. Mean Diff. (Sore Throat - Control)	Statistical Significance: ANOVA p Value
1	6.31-6.33	m	Lipid oxidation product$*$	+	0.030
	0.60-0.62	Broad	Proteins	+	ns
	1.36-1.40	d	Acetoin-CH$_3$	+	ns
	3.44-3.61	s	Glycine-α-CH$_2$	+	0.049
	6.83-6.88	Broad/d	Protein Tyrosine Residues - Unknown multiplet$*$	+	0.012
Sb	ISB (ppm)	^{1}H NMR Resonance Mult.	Metabolic Assignment	Sign of Class. Mean Diff. (Sore Throat - Control)	Statistical Significance: ANOVA p Value
2	2.22-2.27	t	5-Aminovalerate-α-CH$_2$ $**$	+	0.013
	2.29-2.31	Weak m	γ-Aminobutyrate-α-CH$_2$$*$/Propionylglycine-$\alpha$-CH$_2$$*$	-	0.015
	2.78-2.83	m	Aspartate-β-CH$_2$	+	ns
	5.66-5.69	m	Senecioate-α-CH vinylic proton$*$	+	0.026
Sb	ISB (ppm)	^{1}H NMR Resonance Mult.	Metabolic Assignment	Sign of Class. Mean Diff. (Sore Throat - Control)	Statistical Significance: ANOVA p Value
3	2.22-2.27	t	5-Aminovalerate-α-CH$_2$ $**$	+	0.015
	2.78-2.83	m	Aspartate-β-CH$_2$	+	ns
	5.66-5.69	m	Senecioate-α-CH vinylic proton$*$	+	0.017
Sb	ISB (ppm)	^{1}H NMR Resonance Mult.	Metabolic Assignment	Sign of Class. Mean Diff. (Sore Throat - Control)	Statistical Significance: ANOVA p Value
4	2.22-2.27	t	5-Aminovalerate-α-CH$_2$ $**$	+	0.015
	5.66-5.69	m	Senecioate-α-CH vinylic proton$*$	+	0.017
	8.37-8.42	m	1-Methyladenine$*$/Pterin-pyrazine ring proton$*$	+	0.010
Sb	ISB (ppm)	^{1}H NMR Resonance Mult.	Metabolic Assignment	Sign of Class. Mean Diff. (Sore Throat - Control)	Statistical Significance: ANOVA p Value
5	2.22-2.27	t	5-Aminovalerate-α-CH$_2$ $**$	+	0.015
	5.66-5.69	m	Senecioate-α-CH vinylic proton$*$	+	0.017

$*$Tentative assignment (the 6.31-6.33 ppm ISB resonance may arise from a conjugated hydroperoxy- or hydroxydiene lipid oxidation product, and the 6.83-6.88 ppm multipet may arise from 3,4-dihydroxymandelate, 4-hydroxyphenylacetate, pyrocatechol or 3-hydroxymandelate); $**$ For a small number of samples, this ISB also contained an acetone-CH3 group signal (s, δ=2.245 ppm). Abbreviations: ISB, 'Intelligently-Selected' Bucket; s, singlet; d, doublet; t, triplet; m, multiplet; ns, not significant via mixed model ANOVA analysis; Mult, multiplicity.

of the human salivary supernatant specimens. Indeed, the 2.24 ppm resonance was found to be clearly linked to those at δ=1.66 (two sets of overlapping *tt* multiplets) and 3.025 ppm (triplet) of relative intensities 2.0 and 1.0 respectively to that of the 2.24 ppm signal; these signals are ascribable to 5-aminovalerate's 3-/4- and 5-position methylene group protons, with the 2.24 ppm one assigned to the 1-position (α-CH$_2$) ones. With the exception of the 2.29-2.31 ppm spectral bucket, all of the ISBs selected as important disease-determining predictor variables were of a higher salivary concentration in the active sore throat disease class of patients than those in the healthy age-matched control group, and this may partially arise from dehydration, which is a common feature associated with this condition.

The clinical and metabolomic significance of the biomolecular features selected via application of the FuGePSD technique employed here will be reported and discussed in detail elsewhere. However, it should be noted that 5-aminovalerate, one of the key biomarkers detected, is a microbial metabolite generated by oral microflora via a mechanism involving the bacterial catabolism of L-lysine (Fothergill and Guest, 1977) (although it may also be formed endogenously (Callery and Geelhaar, 1984)). Therefore, its elevated salivary concentration in patients with an acute sore-throat condition may reflect an enhanced (localised) level of microbial growth and preponderance in those afflicted. Moreover, acetoin was also found to be upregulated in the salivary metabolome of subjects with an acute sore-throat condition, and this agent is generated via fermentation processes; indeed, it is a catabolite of the butanediol cycle in microorganisms.

Acknowledgment

This work was partially 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.

This paper was written in the Leicester School of Pharmacy from De Montfort University in Leicester, during the research visit of C.J. Carmona. Thanks to Prof. M. Grootveld and Prof. D. Elizondo for the invitation. They are very grateful to De Montfort University, Leicester, UK for the provision of a HEIF collaborative award to support this project.

References

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

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F., 2011. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing 17, 255–287.

Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M., Ventura, S., Garrell, J., Otero, J., Romero, C., Bacardit, J., Rivas, V., Fernández, J., Herrera, F., 2009. KEEL: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems. Soft Computing 13, 307–318.

Atzmueller, M., Puppe, F., 2006. SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery, in: Proc. of the 17th European Conference on Machine Learning and 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, Springer. pp. 6–17.

Atzmueller, M., Puppe, F., Buscher, H.P., 2004. Towards Knowledge-Intensive Subgroup Discovery, in: Proc. of the Lernen - Wissensentdeckung - Adaptivität - Fachgruppe Maschinelles Lernen, pp. 111–117.

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

Bay, S.D., Pazzani, M.J., 2001. Detecting group differences: Mining contrast sets. Data Mining and Knowledge Discovery 5, 213–246.

Berlanga, F.J., del Jesus, M.J., González, P., Herrera, F., Mesonero, M., 2006. Multiobjective Evolutionary Induction of Subgroup Discovery Fuzzy Rules: A Case Study in Marketing, in: Proc. of the 6th Industrial Conference on Data Mining, Springer. pp. 337–349.

Bisno, A.L., Gerber, M.A., Gwaltney, J.M., Kaplan, E.L., Schwartz, R.H., 2002. Infectious Diseases Society of America. Practice guidelines for the diagnosis and management of group A streptococcal pharyngitis. Clinical Infectious Diseases 35, 113,125.

Blackman, R., Tukey, J.W., 1958. The measurement of power spectra : from the point of view of communications engineering. Dover Publications.

Brin, S., Motwani, R., Ullman, J.D., Tsur, S., 1997. Dynamic Itemset Counting and Implication Rules for Market Basket Data, in: Proc. of the 1997 ACM SIGMOD International Conference on Management of Data, ACM Press. pp. 255–264.

Callery, P.S., Geelhaar, L.A., 1984. Biosynthesis of 5-aminopentanoic acid and 2-piperidone from cadaverine and 1-piperideine in the mouse. Journal of Neurochemistry 43, 1631–1634.

Carmona, C.J., Chrysostomou, C., Seker, H., del Jesus, M.J., 2013a. Fuzzy Rules for Describing Subgroups from Influenza A Virus Using a Multi-objective Evolutionary Algorithm. Applied Soft Computing 13, 3439–3448.

Carmona, C.J., González, P., García-Domingo, B., del Jesus, M.J., Aguilera, J., 2013b. MEFES: An evolutionary proposal for the detection of exceptions in subgroup discovery. An application to Concentrating Photovoltaic Technology. Knowledge-Based Systems 54, 73–85.

Carmona, C.J., González, P., del Jesus, M.J., Herrera, F., 2010a. NMEEF-SD: Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery. IEEE Transactions on Fuzzy Systems 18, 958–970.

Carmona, C.J., González, P., del Jesus, M.J., Navío, M., Jiménez, L., 2011a. Evolutionary Fuzzy Rule Extraction for Subgroup Discovery in a Psychiatric Emergency Department. Soft Computing 15, 2435–2448.

Carmona, C.J., González, P., del Jesus, M.J., Romero, C., Ventura, S., 2010b. Evolutionary algorithms for subgroup discovery applied to e-learning data, in: Proc. of the IEEE International Education Engineering, pp. 983–990.

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

Carmona, C.J., Ramírez-Gallego, S., Torres, F., Bernal, E., del Jesus, M.J., García, S., 2012. Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. Expert Systems with Applications 39, 11243–11249.

Carmona, C.J., Ruiz-Rodado, V., del Jesus, M.J., Weber, A., Grootveld, M., González, P., Elizondo, D., 2015. A fuzzy genetic programming-based algorithm for subgroup discovery and the application to one problem of pathogenesis of acute sore throat conditions in humans. Information Sciences 298, 180–197.

Chrysostomou, C., Seker, H., Aydin, N., 2011. 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). pp. 4955–4958.

Chrysostomou, C., Seker, H., Aydin, N., Haris, P., 2010. Complex Resonant Recognition Model in Analysing Influenza A Virus Subtype Protein Sequences, in: 10th IEEE International Conference on Information Technology and Applications in Biomedicine.

Claxson, A., Grootveld, M., Chander, C., Earl, J., Haycock, P., Mantle, M., Williams, S.R., Silwood, C.J.L., Blake, D.R., 1999. Examination of the metabolic status of rat air pouch inflammatory exudate by high field proton NMR spectroscopy. Biochimica et Biophysica Acta-Molecular Basis of Disease 1454, 57–70.

Cordón, O., Herrera, F., Hoffmann, F., Magdalena, L., 2001. Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases. World Scientific.

Cosic, I., 1994. Macromolecular bioactivity: is it resonant interaction between macromolecules: Theory and applications. IEEE transactions on bio-medical engineering 41, 1101–1114.

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

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

Dong, G.Z., Li, J.Y., 2005. Mining border descriptions of emerging patterns from dataset pairs. Knowledge and Information Systems 8, 178–202.

Eiben, A.E., Smith, J.E., 2003. Introduction to evolutionary computation. Springer.

Fothergill, J.C., Guest, J.R., 1977. Catabolism of l-lysine by Pseudomonas aeruginosa. Journal of general microbiology 99, 139–145.

Gamberger, D., Lavrac, N., 2002. Expert-Guided Subgroup Discovery: Methodology and Application. Journal Artificial Intelligence Research 17, 501–527.

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

Goldberg, D.E., 1989. Genetic Algorithms in search, optimization and machine learning. Addison-Wesley Longman Publishing Co., Inc.

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

Grootveld, M., Atherton, M.D., Sheerin, A.N., Hawkes, J., Blake, D.R., Richens, T.E., Silwook, C.J.L., Lynch, E., Claxson, A.W.D., 1998. In vivo absorption, metabolism, and urinary excretion of alpha,beta-unsaturated aldehydes in experimental animals. Relevance to the development of cardiovascular diseases by the dietary ingestion of thermally stressed polyunsaturate-rich culinary oils. The Journal of Clinical Investigation 101, 1210–1218.

Grootveld, M., Sheerin, A., Atherton, M., Millar, A.D., Lynch, E.J., Blake, D.R., Naughton, D.P., 1996. Biomedical applications of NMR Spectroscopy. John Wiley and Sons. volume 25. chapter Applications of high resolution NMR analysis to the study of inflammatory diseases at the molecular level. pp. 295–327.

Grosskreutz, H., Rueping, S., Wrobel, S., 2008. Tight optimistic estimates for fast subgroup discovery, in: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, pp. 440–456.

Henry, R.F., Graefe, P.W.U., 1971. Zero padding as a means of improving definition of computed spectra. Published for Environment Canada by Dept. of Energy, Mines and Resources, Marine Sciences Branch.

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

Herrera, F., Carmona, C.J., González, P., del Jesus, M.J., 2011. An overview on Subgroup Discovery: Foundations and

Applications. Knowledge and Information Systems 29, 495–525.

Holland, J.H., 1975. Adaptation in natural and artificial systems. University of Michigan Press .

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

del Jesus, M.J., González, P., Herrera, F., Mesonero, M., 2007b. Evolutionary Fuzzy Rule Induction Process for Subgroup Discovery: A case study in marketing. IEEE Transactions on Fuzzy Systems 15, 578–592.

Kloesgen, W., 1996. Explora: A Multipattern and Multistrategy Discovery Assistant, in: Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence, pp. 249–271.

Kloesgen, W., Zytkow, J., 2002. Handbook of Data Mining and Knowledge Discovery. Oxford.

Konar, A., 2005. Computational Intelligence: Principles, Techniques and Applications. Springer-Verlag New York, Inc.

Kralj-Novak, P., Lavrac, N., Webb, G.I., 2009. Supervised Descriptive Rule Discovery: A Unifying Survey of Constrast Set, Emerging Pateern and Subgroup Mining. Journal of Machine Learning Research 10, 377–403.

Kuncheva, L., 2000. Fuzzy classifier design. Springer.

Lavrac, N., Cestnik, B., Gamberger, D., Flach, P.A., 2004a. Decision Support Through Subgroup Discovery: Three Case Studies and the Lessons Learned. Machine Learning 57, 115–143.

Lavrac, N., Flach, P.A., Zupan, B., 1999. Rule Evaluation Measures: A Unifying View, in: Proc. of the 9th International Workshop on Inductive Logic Programming, Springer. pp. 174–185.

Lavrac, N., Kavsek, B., Flach, P.A., Todorovski, L., 2004b. Subgroup Discovery with CN2-SD. Journal of Machine Learning Research 5, 153–188.

Lemanska, A., Grootveld, M., Silwood, C.J.L., Brereton, R.G., 2011. Chemometric variance analysis of 1H NMR metabolomics data on the effects of oral rinse on saliva. Metabolomics 8, 64–80.

Morens, D.M. and Taubenberger, J.K. and Fauci, A.S., 2009. The persistent legacy of the 1918 influenza virus. The New England journal of medicine 361, 225.

Moscona, A., 2005. Neuraminidase inhibitors for influenza. New England Journal of Medicine 353, 1363.

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

Noda, E., Freitas, A.A., Lopes, H.S., 1999. Discovering interesting prediction rules wih a genetic algorithm. IEEE Congress on Evolutionary Computation 2, 1322–1329.

Palm, R., Hellendoorn, H., Driankov, D., 1997. Model Based Fuzzy Control. Springer.

Pedrycz, W., 1996. Fuzzy Modelling: Paradigms and Practices. Kluwer Academic Publishers.

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

Rojas, R., 1996. Neural Networks: A Systematic Introduction. Springer-Verlag New York, Inc.

Siebes, A., 1995. Data Surveying: Foundations of an Inductive Query Language, in: Proc. of the 1st International Conference on Knowledge Discovery and Data Mining, AAAI Press. pp. 269–274.

Silwood, C.J.L., Lynch, E., Claxson, A.W.D., Grootveld, M., 2002. 1H and 13C NMR spectroscopic analysis of human saliva. Journal of Dental Research 81, 422–427.

Sundararaja., D., 2001. The Discrete Fourier Transform: Theory, Algorithms and Applications. World Scientific.

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

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

Wongravee, K., Lloyd, G.R., Silwood, C.J.L., Grootveld, M., Brereton, R.G., 2010. Supervised Self Organizing Maps (SOMs) for classification and variable selection: illustrated by application to NMR metabolomic profiling. Analytical Chemistry 82, 628–638.

Worrall, G., 2006. Theres a lot of it about: acute respiratory infection in primary care. Abingdon Engl: Radcliffe Publishing Ltd. chapter Acute sore throat. pp. 24–36.

Wrobel, S., 1997. An Algorithm for Multi-relational Discovery of Subgroups, in: Proc. of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery, Springer. pp. 78–87.

Wrobel, S., 2001. Inductive logic programming for knowledge discovery in databases. Springer. chapter Relational Data Mining. pp. 74–101.

Yager, R.R., Filev, D.P., 1994. Essentials of Fuzzy Modeling and Control. John Wiley & Sons, Inc.. 1st edition.

Zadeh, L.A., 1965. Fuzzy sets. Information Control 8, 338–353.

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

Zadeh, L.A., 1994. Soft Computing and Fuzzy Logic. IEEE Software 11, 48–56.

Zitzler, E., Laumanns, M., Thiele, L., 2002. SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimization, in: International Congress on Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems, pp. 95–100.