Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Detection of Safe and Harmful Bioaerosols by Means of Fuzzy Classifiers ⋆

Pietari Pulkkinen * **Jarmo Hytönen** ** **Hannu Koivisto** *

* *Tampere University of Technology, Department of Automation Science and Engineering, P.O. Box 692, FIN-33101 Tampere Finland (Tel: +358 3 3115 2655; e-mail: firstname.lastname@tut.fi).*
** *Dekati Ltd. Osuusmyllynkatu 13 FIN-33700 Tampere Finland (e-mail: jarmo.hytonen@dekati.fi)*

Abstract: This paper aims to create a fuzzy classifier (FC) to be used in a recently developed bioaerosol detector. The main requirements for FC are high true positive (TP) rate, low false positive (FP) rate, and interpretability, which is measured by transparency of fuzzy partition. Due to the contradicting nature of the above requirements, FCs are identified by hybrid genetic fuzzy system (GFS), which initializes the population using decision trees (DTs) and simplification operations. Then, a multiobjective evolutionary algorithm (MOEA) is run in order to find a Pareto-optimal set of FCs. During MOEA optimization, heuristic rule and rule condition removal is applied to keep the rule base consistent. Real-world bioaerosol data, collected from Umeå trial field, Sweden, and from laboratory of Finnish Defense Forces Technical Research Center, were used to validate the proposed GFS. By means of it, a widely spread set of interpretable and accurate FCs was obtained. Moreover, an FC based on this project was installed into the bioaerosol detector and the preliminary tests proved its capability in distinguishing between safe and harmful bioaerosols.

Keywords: Bioaerosol Detector; Fuzzy Classifiers (FCs); Multiobjective Evolutionary Algorithms (MOEAs); Genetic Fuzzy System (GFS); Interpretability.

1. INTRODUCTION

The purpose of a bioaerosol detector is to distinguish between safe and harmful bioaerosols according to several measurement signals. The commonly applied signals are UV-fluorescent and the size of the particles Sivaprakasam et al. (2004), Janka et al. (2007) and the first bioaerosol alarm system based on those signals was proposed in 1997 Hairston et al. (1997). There are, however, some aerosols (e.g. soot from diesel engines) which may cause fluorescent response. Hence, those aforementioned signals are not sufficient for reliable detection. In this paper, therefore, the UV-fluorescence detection optics are combined with a special background-aerosol detector system proposed by Janka et al. (2007). That increases the reliability by reducing false alarms.

The goal of this work is to identify a model, which reasons based on the real-time measurements whether there are harmful or safe bioaerosols in the air. The further analysis is done in laboratory after an alarm is issued. That can be an expensive and time consuming operation Hairston et al. (1997). Hence, it is important to minimize the number of false alarms. Furthermore, the bioaerosol data collected from the field and laboratory are highly imbalanced; there are much more data points representing harmless particles than harmful particles. Therefore, true positive (TP) rate and false positive (FP) rate are used as a accuracy metrics instead of commonly used misclassification rate.

Understanding the reasoning of the model builds up the confidence that the model actually works reasonably, which in real-world problems, is one of the crucial requirements for the model Elder and Pregibon (1996). Therefore, the model is required to be as interpretable as possible.

Fuzzy classifiers (FCs) can be highly interpretable due to their linguistic rules. Furthermore they can be very accurate. However, data-driven methods often lead into more complex models than necessary and therefore interpretability is lost Setnes et al. (1998). Recently, more FCs are identified by using evolutionary algorithms (EAs), because of their good learning capabilities for complex problems. Those approaches are often called genetic fuzzy systems (GFS) Cordón et al. (2004). They may have multiple objectives, for example number of rules, total number of conditions and misclassification rate (c.f. Ishibuchi et al. (2001)), which are to be minimized simultaneously. The GFS can converge into a single solution when aggregated fitness function is used or into a set of Pareto-optimal solutions when multi-objective evolutionary algorithms (MOEAs) are used.

Commonly the initial population for GFS is created randomly or manually Ishibuchi et al. (2006), Setzkorn and Paton (2005), Gómez-Skarmeta et al. (1998), while better convergence due to reduction of the search space is obtained by adequate initialization Haubelt et al. (2005),

⋆ This project was funded by Finnish defence forces chemical, biological, nuclear, status (CBNS) technology program.

Poles et al. (2006). That can be done, for example, by using DTs or clustering algorithms and transforming DTs or clusters into FMs Roubos and Setnes (2001), Abonyi et al. (2003), Pulkkinen and Koivisto (2007b).

In this work we adopt hybrid GFS, which initializes the population using DT algorithm and simplification operations. Then, the initial population is further optimized by MOEA. Hybrid GFS is a refinement of our previous work Pulkkinen and Koivisto (2007b) and its objectives are transparency of fuzzy partition and TP and FP rates. When transparency of fuzzy partition is used as an objective, intuitive linguistic values for linguistic variables are obtained. Moreover, inconsistencies in rule base are avoided by heuristic rule and rule condition reduction.

The obtained results confirm the usefulness of the proposed hybrid GFS. By means of it, a widely spread set of interpretable and accurate FCs was obtained. In January 2007, an FC based on this project was installed into the bioaerosol detector and the preliminary tests proved its capability in distinguishing between safe and harmful bioaerosols.

This paper is organized as follows. Section 2 introduces the constructed bioaerosol detector and discusses the data collection. Then, section 3 introduces the proposed hybrid GFS. After that in section 4 the obtained results are presented. Finally, conclusions are given in section 5.

2. BIOAEROSOL DETECTOR

Fig. 1 shows the schematic diagram of bioaerosol detector proposed by Janka et al. (2007). To reduce the false alarms caused by non-bioaerosols it includes not only UV-fluorescence and particle size signals but also a special background-aerosol detector system, which makes it different from the conventional bioaerosol detectors.

In a nutshell, the bioaerosol detector works as follows. First, the particles go through a size selective sampling to a concentrator. In order to prevent the pollens to enter, the cut-off size for this inlet is selected as $< 7\mu m$. The concentrator then gains the larger particles ($> 2\mu m$) by factor > 500 and they go to the optical measurement unit (also called primary unit) where their UV-fluorescence and elastic scatter are measured. The rest of the particles (e.g. the smaller ones $< 2\mu m$) go to the secondary unit. Because the particles from combustion processes (e.g. exhaust gases from diesel engine) are of this size, the secondary unit is in key role in preventing false alarms Janka et al. (2007).

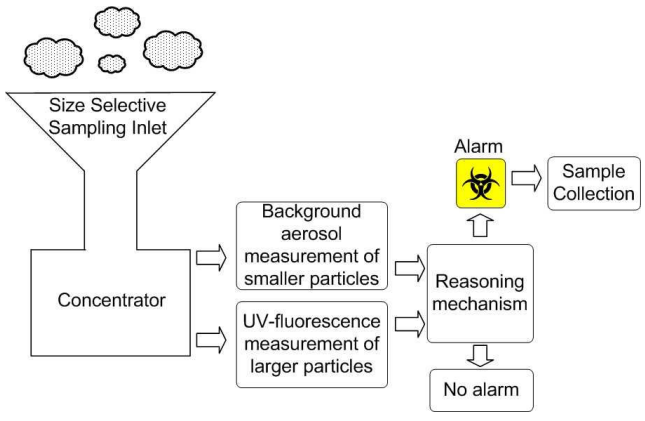

Fig. 1. A schemata of bioaerosol detector.

When an alarm is raised, the sample collection is automatically started by channeling the aerosol flow to a dry filter. Then the filter is removed and brought to the laboratory for further investigation Janka et al. (2007).

2.1 Collecting and Preprocessing of the Data

The real-world bioaerosol data were collected both from the field and from the laboratory. The field measurement campaign was performed at Umeå trial field in Sweden during early autumn 2006. The laboratory measurements were performed at Finnish Defense Forces Technical Research Center at the end of October 2006.

One of the major challenges for preparing the data for supervised learning algorithms is to assign each input data point to adequate class label. In this work there are only two classes, namely alarm and normal. Of course during the measurement period, it is exactly known when particles are spread. Furthermore, it is known whether the distributable particles are harmful or safe. However, there are many challenges in assigning the class label for the data, especially when the field measurements are considered. Some of them are listed next:

- It is difficult to estimate when particles have spread from distribution point to the measurement point. That depends on several factors like direction and velocity of the wind.
- There can be some disturbances, for instance, dust caused by walking people and smoke caused by cars or cigarettes.

In this work, the output was labeled based on the field log and expert knowledge. It was roughly known when the particles should reach the measurement system. Then, when a rise in a certain measurement value occurred, it was marked as the point particles has reached the measurement system. If those particles were harmful, the output was marked as alarm; otherwise it was marked as normal. Then when the value of that measurement descended, it was marked as the point the spreading of the particles was finished. Before and after the spreading period the output was marked as normal.

However, between the start and the end of the spreading period the direction of the wind may change. Therefore, between the start and the end of the spreading there may be periods during which no particles are reached to the measurement system. If those data points are marked as alarm, that causes the data to become inconsistent and therefore causes difficulties for the supervised learning algorithms. Thus, a threshold for the aforementioned measurement was defined based on expert knowledge. Therefore, even harmful particles are spread, but if the threshold is not exceeded those data points are still marked as normal. From the data points not exceeding the threshold, only a presentative subset was chosen in order to reduce the computational costs.

2.2 Anomalies in the Data

When large amounts of real-world data are collected, the collected data contain anomalies. Furthermore, the measurement system is currently in prototype phase and many changes and improvements are expected in the near future. Because of that, it is natural that the collected data contain anomalies and extensive noise, which will not be the case anymore when the system is more established.

The collected data have some anomalies in absolute values of the measurements; there are changes in the values of the measurements when there is nothing spread in the air. Naturally those values, which are called in this work *zero values*, should always be approximately the same, but in this case quite significant variations exist. Moreover, one of the measurements was very sensitive to temperature and caused trends in the data.

That makes the development of accurate models a very difficult task. However, it is important for the continuation of the project to show that the developed bioaerosol detector has potential for distinguishing between harmful and harmless particles. Keeping those factors in mind, it is not meaningful to clean the data from all anomalies. Anyway, when the system is evaluated online the obtained measurement data may currently have some anomalies and the model should still work. Thus, only the data points with most significant errors, for example a negative measurement value for a measurement which should always be positive or very strong trends, were removed. Also some zero values were slightly modified.

3. PROPOSED HYBRID GFS

The proposed hybrid genetic fuzzy system (GFS) is presented in Fig. 2. First, a fuzzy classifier (FC) is created by a crisp decision tree (DT) algorithm. That is clearly a better starting point for further optimization than commonly applied random initialization. However, due to crispness of DT and the noise in real-world data, this FC is overly complex and can be simplified Abonyi et al. (2003). Therefore, the initial FC goes through merging of similar fuzzy sets, which may lead to similar Setnes et al. (1998) or inconsistent rules. Those rules are heuristically removed in order to improve the convergence of MOEAs and to reduce computational costs of fitness evaluations. After that, the rest of the population is created by modifying the simplified FC, such that, the initial population is widely spread. That is beneficial to the convergence of MOEAs Haubelt et al. (2005), Poles et al. (2006). Finally, a MOEA is applied to find a set of widely spread Pareto-optimal FCs. During MOEA optimization the offspring population goes through rules and rule conditions reduction in order to prevent the rule base having inconsistent rules.

As DT algorithm, C4.5 Quinlan (1993) is applied. Its advances include selection of input variables and partition of input space with non-fixed number of hyper-rectangles Abonyi et al. (2003); Pulkkinen and Koivisto (2007a). It is a well known and widely used algorithm and therefore no details are given in this paper. A reader may refer to Quinlan (1993); Abonyi et al. (2003) for further details.

As MOEA component, NSGA-II Deb et al. (2002), a popular and commonly applied MOEA is used. Its strengths include an efficient method for constraint-handling , a fast non-dominated sorting procedure and a parameterless crowding distance measure for maintaining diversity of population. Its details are not presented in this paper, since it is well documented in Deb et al. (2002).

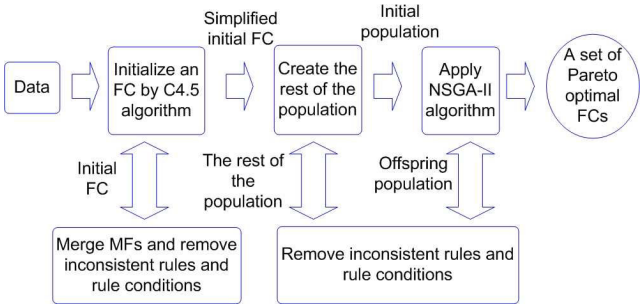

Fig. 2. Proposed hybrid genetic fuzzy system.

The rest of this section is organized as follows. First FCs are briefly presented. Then, the proposed hybrid GFS, which is a refinement of our earlier work Pulkkinen and Koivisto (2007b), is described in detail using the same notations as in our earlier work. Then, metrics for accuracy and interpretability are discussed. Finally, the fitness function applied in this paper is presented.

3.1 Fuzzy Classifiers

A fuzzy classification rule consists of fuzzy sets in the antecedent and a class label in the consequent. Let us denote the data set with D data points and n variables as $\mathbf{Z} = [\mathbf{X}\ \mathbf{y}]$, where input matrix $\mathbf{X}$ and output vector $\mathbf{y}$ are given as:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{D,1} & x_{D,2} & \dots & x_{D,n} \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_D \end{bmatrix}. \tag{1}$$

According to Abonyi et al. (2003) fuzzy classification can be performed as follows:

$$Rule_i: \text{ If } x_1 \text{ is } A_{i,1}, \dots, \text{ and } x_n \text{ is } A_{i,n} \text{ then } g_i, \quad i = 1, \dots, R, \tag{2}$$

where R is the number of rules, $A_{i,j}, j = 1, \dots, n$ is a membership function, $g_i \in \{1, \dots, C\}$ is the rule consequent and C is the number of different classes in data set. For each data point $\mathbf{x}_k$, the degree of fulfillment of a rule is computed as:

$$\beta_i(\mathbf{x}_k) = \prod_{j=1}^{n} A_{i,j}(x_{k,j}). \tag{3}$$

The rule with the highest degree of fulfillment β^* is declared as winner rule (i.e. Winner takes all strategy). The output of the classifier is the rule consequent associated to that rule.

3.2 Initialization of FCs

First a DT is created by C4.5 algorithm and converted into an FC like presented in Abonyi et al. (2003). That can be

done without decomposition error, if trapezoidal membership functions (MFs) are applied. However, application of generalized bell (gbell) MFs have several benefits. Gbell MFs may have better fit to the data Setnes and Roubos (2000) and they have three parameters in contrast to four parameters of trapezoidal MFs. Furthermore, since the parameters of gbell MFs can be optimized independently, standard mutation and crossover operators of MOEAs can be used without any feasibility check of MFs parameters. Because of the above reasons gbell MFs are applied in this paper and they are defined as:

$$\mu(x; a, b, c) = \frac{1}{1 + \left|\frac{x-c}{a}\right|^{2b}}, \quad (4)$$

where x is the data point, and a, b and c are the parameters of a gbell MF. The value of a defines the width of an MF. In this paper it is required that a should be at least 1% of variable range $\chi = ubound - lbound$, where $ubound$ and $lbound$ are respectively the upper and lower bounds of a variable. In addition, no MFs should be wider than χ and therefore $0.005\chi < a < \chi/2$. The value of b defines the fuzziness of an MF. If it is set to a high value, an MF is almost a crisp function. Moreover, if $b \approx 0$ an MF can cover large areas of universe of discourse, therefore leading to covering of fuzzy sets, which will be illustrated later in subsection 3.5. Therefore, $1 < b < 10$. Center of an MF should be inside the variable range. Thus, $lbound < c < ubound$.

Simplification of Initial FC The initial FC is commonly overly complex due to the axis parallel partition of crisp DT Abonyi et al. (2003) and the noise in real-world data. Highly similar fuzzy sets of the initial FC are merged according to Setnes et al. (1998), Wang et al. (2005):

$$S(A_i, A_j) = \frac{|A_i \bigcap A_j|}{|A_i \bigcup A_j|} \approx \frac{\sum_{k=1}^{p}[\mu_i(x_k) \wedge \mu_j(x_k)]}{\sum_{k=1}^{p}[\mu_i(x_k) \vee \mu_j(x_k)]}, \quad (5)$$

where $\bigcap$ and $\bigcup$ are the set theoretic intersection and union, respectively. Minimum is marked by $\wedge$ and maximum by $\vee$. The left hand side of the formula is commonly approximated by calculating p membership values μ for fuzzy sets i and j in discrete universe $U = \{x_j | j = 1, 2, \ldots, p\}$ Wang et al. (2005). All pairs exceeding the user specified threshold Δ are merged.

The parameters of the fuzzy set $A^{'}$, which replaces fuzzy sets i and j are:

$$a' = \frac{\max(c_i + a_i, c_j + a_j) - \min(c_i - a_i, c_j - a_j)}{2}, \quad (6)$$

$$b' = \frac{b_i + b_j}{2} \text{ and } c' = \frac{c_i + c_j}{2}. \quad (7)$$

By result of merging, the rule base may have similar Setnes et al. (1998) or inconsistent rules and rule conditions. It is beneficial to remove them before creating the rest of the chromosomes sharing the same structure (i.e. the same number of possible rules, the same total number of possible MFs and the same number of possible input variables) with simplified initial FC. Heuristic rule removal is applied for that purpose and will be discussed later in subsection 3.3. Merging of fuzzy sets is only applied to the initial FC created by DT.

Structure of a Chromosome The simplified initial FC is the first member (chromosome) of the initial population. Its structure is coded, such that, it can be optimized using standard MOEA, such as NSGA-II. The rest $N-1$ chromosomes, where N is the population size, share the same structure with the simplified initial FC.

The structure of an FC includes antecedents of the rules $\mathbf{A}$ and parameters of fuzzy sets $\mathbf{P}$. $\mathbf{A}$ is given as:

$$\mathbf{A} = A_{i,j},\ i = 1, \ldots, R,\ j = 1, \ldots, n_s, \quad (8)$$

where R denotes the number of rules in simplified initial FC and n_s stands for the number of variables selected from n variables initialization phase. Naturally $n_s \leq n$, but usually $n_s < n$. $A_{i,j} = \{0, 1, \ldots, M_j\}$ indicates which MF is used for variable j in rule i and M_j is the number of MFs assigned to variable j in simplified initial FC. If variable j is not used in rule i, then $A_{i,j} = 0$. If rule i is not used in an FC, then $\forall j, A_{i,j} = 0$. However, for simplified initial FC there is no rule i, for which $\forall j, A_{i,j} = 0$.

Parameter vector $\mathbf{P}$ is presented as:

$$\mathbf{P} = P_{l,k},\ l = 1, \ldots, \gamma,\ k = 1, \ldots, \beta, \quad (9)$$

where γ is the number of parameters used to define an MF and $\beta = \sum_{j=1}^{n_s} M_j$ is the total number of MFs in simplified initial FC. In this paper Gbell MFs are used, so $\gamma = 3$.

Consequent part of the fuzzy rule $\mathbf{g} = [g_1, \ldots, g_R]$ is not included into an individual. It is static and created in initialization phase by DT and by simplification operators. So, MOEA is used to select rules, rule antecedents and parameters of MFs for the pre-specified class labels. The total number of parameters θ to be optimized by MOEA is therefore given as:

$$\theta = R \times n_s + \gamma \times \beta. \quad (10)$$

Each parameter is restricted with lower and upper bounds defined in current and previous subsections. Therefore the number of constrains is $2 \times \theta$.

Initialization of the Rest of The Population The rest $N-1$ individuals of the population, which is of size N, are created by randomly replacing some parameters of the simplified initial FC. The replacement algorithm creates a set of widely distributed chromosomes as follows:

Repeat for $I = 1, \ldots, N-1$, where I is the chromosome iterator.
Step 1: Compute the number of replaceable parameters m:

$$m = \text{round}\left(\frac{I}{(N-1)} \times \theta\right), \quad (11)$$

where *round* stands for the operator rounding the result to the nearest integer.
Step 2: Choose randomly m parameters out of θ.
Step 3: Replace them by randomly generating m parameters between their corresponding limits.
End for

So a population of widely distributed chromosomes is created. They all share the same structure with the simplified initial FC. The rule base of the rest $N-1$ chromosomes may contain inconsistencies due to the random

replacement algorithm, so they go through heuristic rule reduction presented next.

3.3 Heuristic Rule and Rule Condition Reduction

It is beneficial to remove rules and rule conditions heuristically in order to guide and speed up the evolutionary search. The following heuristics are applied in this paper:

(1) If there are rules with exactly the same antecedent part, all but one of them are removed Setnes et al. (1998). The preserved rule is randomly selected.
(2) There can be rules of different length in which all conditions of the shorter rule(s) are present in the longer rule(s). Those rules are inconsistent. The longer rule(s) will never obtain higher degree of firing than the shorter rule(s), because the T-norm in this paper is product. Out of those inconsistent rules only one rule is preserved. By uniform chance, the preserved rule is either the longest rule (i.e. the most specific rule) or it is randomly selected out of the inconsistent rules.
(3) If there are conditions, which are present in all of the rules, they are removed from all of them Pulkkinen and Koivisto (2007a).

These heuristics are applied to the whole initial population. Furthermore, during MOEA optimization whole offspring population goes through heuristic rule and rule condition reduction. Therefore, there is no inconsistencies in the rule base in any of the chromosomes.

When the initial FC generated by DT is simplified by merging of fuzzy sets and by removing rules and rule conditions, the number of optimized parameters θ is decreased. But since the whole population shares the same structure, θ is not affected when the other FCs (i.e. the $N-1$ FCs of initial population and the whole offspring population) go through heuristic rule and rule condition reduction. Then simply the removed conditions are set to 0.

3.4 Metric for Classifier's Performance

It is well known that accuracy (i.e. the proportion of correctly classified data points to the total number of data points) is not an optimal metric for classifier's performance when misclassification costs and/or class distributions are not known Provost et al. (1998); Fawcett (2001); Setzkorn and Paton (2005); Ben-David (2007). In this work, neither misclassification costs nor class distributions are even. The preprocessed data consist of 80.72% of data points with class label normal, and only 19.28% of data points with class label alarm [1]. Thus, true positive (TP) and false positive (FP) rates Fawcett (2001) are used as accuracy metrics:

$$\text{TP rate} = \frac{\text{positives correctly classified}}{\text{total positives}}, \tag{12}$$

$$\text{FP rate} = \frac{\text{negatives incorrectly classified}}{\text{total negatives}}, \tag{13}$$

where positives and negatives are respectively the datapoints labeled as alarms and normal states.

[1] Therefore the accuracy of majority class classifier would 80.72%.

3.5 Interpretability of FCs

Often interpretability of FCs is measured by calculating the number of rules and total number of conditions in rules (total rule length) Ishibuchi et al. (2001); Setzkorn and Paton (2005); Ishibuchi and Nojima (2007). Those metrics, however, does not indicate whether fuzzy partition is transparent or not. Instead of those metrics, slightly modified versions of interpretability metrics proposed in Kim et al. (2006), namely the length of overlap and the length of discontinuity between fuzzy sets, are used. In a nutshell, it is desired that the intersection value of two fuzzy sets would lie between user specified constants α_L and α_H. If the intersection value is higher than α_H, overlap penalty P_{OL} is added, whereas if it is less than α_L, discontinuity penalty P_{DC} is added.

Those penalties, however, issue a very small penalty in cases when a wide fuzzy set covers a narrow fuzzy set (e.g. complete or restricted covering Wang et al. (2005)), albeit the partition is far from transparent. Moreover, in cases of relaxed covering Wang et al. (2005), it is possible that no penalty at all is issued (see Figure 3(c). Also the commonly used similarity measure for fuzzy sets in formula (5) is not informative in cases of covering Wang et al. (2005). Thus, the middle value penalty P_{MV} is introduced to tackle that problem.

Overlap Penalty To calculate P_{OL}, left and right intersection points for the user specified level $\alpha_H = 0.6$ (the same value as in Kim et al. (2006)) are computed first for all fuzzy sets. Since gbell mfs are applied the intersection points can be computed as:

$$I_L(\alpha) = c - a\left(\frac{1-\alpha}{\alpha}\right)^{\frac{1}{2b}}, I_R(\alpha) = c + a\left(\frac{1-\alpha}{\alpha}\right)^{\frac{1}{2b}}. \tag{14}$$

P_{OL} is computed according to Kim et al. (2006):

$$P_{OL} = \frac{1}{n_s}\sum_{i=1}^{n_s}\frac{1}{N_{ov}^i}\sum_{j=1}^{N_{ov}^i}\frac{\lambda_{i,j}}{\chi_i}, \tag{15}$$

where $\lambda_{i,j}$ is the length of jth overlap between two MFs in input variable i. It is computed using the left and right intersection points, like illustrated in Fig. 3(a). N_{ov}^i is the number MF pairs in input variable i, which may overlap:

$$N_{ov}^i = \binom{M_i}{2} = \frac{M_i!}{2(M_i-2)!}, \tag{16}$$

where $M_i > 2$ is the number of active fuzzy sets in input variable i. If there are only 2 MFs, $N_{ov}^i = 1$. P_{OL} is not calculated for a certain variable, if the number of active MFs assigned to it is less than 2.

Discontinuity Penalty Similarly to P_{OL}, computing P_{DC} is started with computing the left and right intersection points for the user specified level $\alpha_L = 0.1$ (the same value as in Kim et al. (2006))). Then, P_{DC} is computed slightly differently than in Kim et al. (2006) as proportion of total length of discontinuity to range χ:

$$P_{DC} = \frac{1}{n_s}\sum_{i=1}^{n_s}\sum_{j=1}^{G_i}\frac{\psi_{i,j}}{\chi_i}, \tag{17}$$

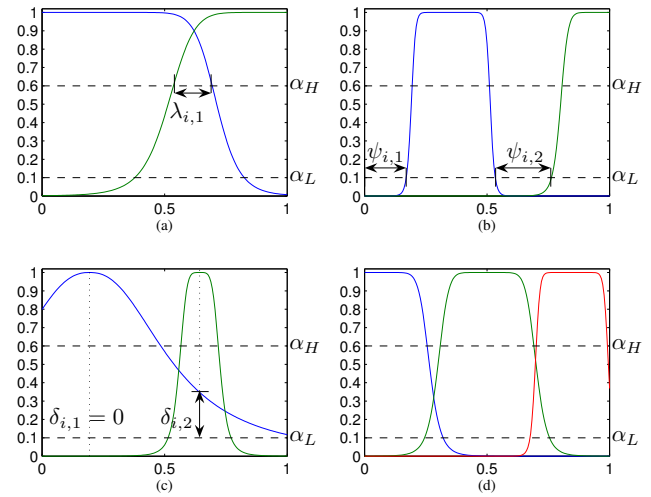

Fig. 3. Four fuzzy partitions: (a) too much overlap, (b) discontinuous partition, (c) relaxed covering, (d) transparent partition.

where G_i is the number of discontinuities and $\psi_{i,j}$ is the length of the jth discontinuity in variable i, respectively (see also Fig. 3(b)). P_{DC} is not calculated for a certain variable, if there are no active MFs assigned to it.

Middle Value Penalty P_{MV} is added to prevent relaxed covering of MFs. In Fig. 3(c) an example of relaxed covering is shown and based on P_{OL} and P_{DC} no penalty is given, even the partition is not transparent. This problem is tackled in this paper by adding P_{MV} if the value of another fuzzy set is higher than α_L in the center of another fuzzy set (see Fig. 3(c)):

$$P_{MV} = \frac{1}{n_s}\sum_{i=1}^{n_s}\delta_i, \tag{18}$$

where

$$\delta_i = \begin{cases} \dfrac{\delta_i^* - \alpha_L}{1-\alpha_L} & \text{if } \delta_i^* > \alpha_L \\ 0 & \text{if } \delta_i^* \leqslant \alpha_L \end{cases}, \tag{19}$$

where δ_i^* is the maximum middle value in variable i:

$$\delta_i^* = \max(\underbrace{\mu_{i,j}(c_k; a_j, b_j, c_j)}_{j\neq k}), \tag{20}$$

where $j = 1, \ldots, M_i, k = 1, \ldots, M_i$.

3.6 Overall Fitness Function

The objectives to be minimized need to be selected carefully in order to avoid deterioration of search efficiency due to increase in the number of objectives Purshouse and Fleming (2003); Hughes (2005). Since TP rate and FP rate are the crucial requirements for the detector to meet they need to be selected as objectives. Transparency of fuzzy partition is another important objective. Therefore transparency penalty T:

$$T = P_{OL} + P_{DC} + P_{MV}, T \in [0, 3) \tag{21}$$

needs to be minimized. It will be shown later in section 4 that by minimizing T, usually the number of rules and rule conditions is also reduced. Thus, they are not selected as objectives like in many other studies (e.g. Ishibuchi et al. (2001); Setzkorn and Paton (2005); Ishibuchi and Nojima (2007)). Hence, the three objectives to be minimized are:

$$o_1 = 1 - \text{TP rate}; o_2 = \text{FP rate}; o_3 = T \tag{22}$$

To avoid impractical solutions, FP and TP rates are constrained inside a square, which its side is d. In this paper it is required that TP rate is at least 0.5 and FP rate the most 0.5. Thus, $d = 0.5$. The normalized constraints are:

$$constraint_1 = \frac{d - o_1}{1 - d} \geq 0; constraint_2 = \frac{d - o_2}{1 - d} \geq 0 \tag{23}$$

Table 1. Parameters used in this paper were exactly the same as in our former study Pulkkinen and Koivisto (2007b). The same crossover and mutation probabilities and distribution indexes were used in Deb et al. (2002).

Distribution index for mutation	20
Distribution index for cross-over	20
Cross-over probability	0.9
Mutation probability	$1/\theta$
Pruning confidence (C4.5 algorithm)	5

4. RESULTS

The collected and pre-processed data contained 10268 data points. Each data point consisted of 4 input variables, labeled for confidentiality reasons as variable A, B, C, and D, and a class label (alarm or normal). Data were divided into train and test sets. The fuzzy classifiers (FCs) were identified using only the train data, which contained 80% of data, and tested with the rest 20% of data.

Initial FC generated by C4.5 was overly complex, containing 47 rules and 172 rule conditions. Its fuzzy partition was very complex and the transparency penalty T was 1.062. Thus, fuzzy sets merging threshold Δ was set to 0.25 in order to reduce the complexity. After merging the similar fuzzy sets and performing heuristic rule and rule condition removal, the number of rules and rule conditions were 21 and 71, respectively. The fuzzy partition was more transparent and T was reduced into 0.755.

Then, the proposed genetic fuzzy system (GFS) was run with population size and number of generations both set to 1000 ($1000 \times 1000 = 10^6$ fitness evaluations). The rest of the parameters are specified in Table 1.

As a result of the run, a widely spread set of FCs was obtained, which is shown in Fig. 4. There were FCs with TP rate as high as 1, FCs with FP rate as low as 0.009 and also highly transparent FCs with transparency penalty 0 (completely transparent partition). However, due to the contradicting nature of the objectives, all of those extreme values were not present in a single FC. During the MOEA optimization the average number of rules was reduced from 15.641 to 5.599 and the average number of rule conditions from 54.286 to 10.495. That happened even though those were not used as fitness objectives, which means that they are somewhat correlated with the average value of T which was reduced from 1.095 to 0.099 during the MOEA optimization.

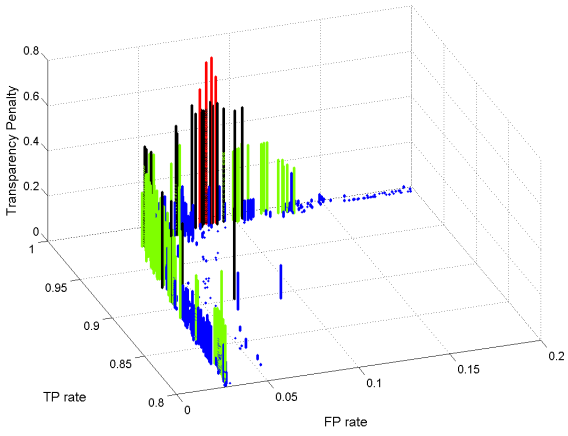

Fig. 4. Obtained FCs for test set. Only the FCs having TP rate at least 0.8 and FP rate the most 0.2 are presented. Colors blue, green, black and red indicate the FCs with transparency penalty of < 0.2, < 0.4, < 0.6, and > 0.6, respectively.

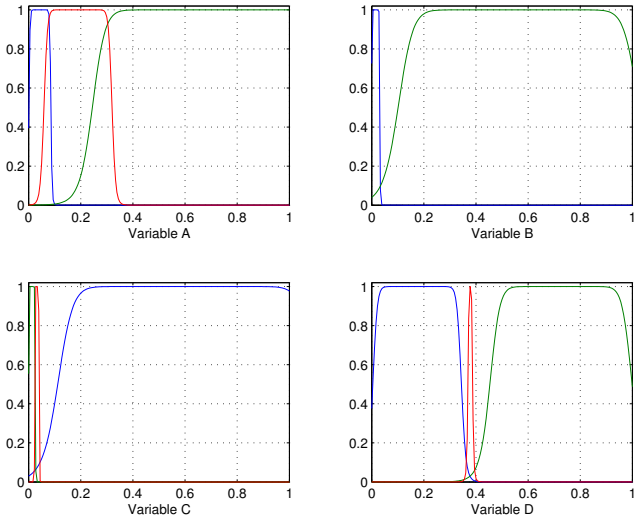

Fig. 5. Fuzzy partition with transparency penalty of 0.014. Partition is highly transparent.

From the obtained solutions a solution based on the preferences can be selected and the rest of the solutions may be stored for possible usage. A user who demands highly transparent fuzzy partition may, for example, select an FC with $TP_{test} = 0.944$, $FP_{test} = 0.046$, and $T = 0.014$, whereas a user demanding higher TP rate may select an FC with $TP_{test} = 0.964$, $FP_{test} = 0.044$, and $T = 0.254$. Fuzzy partitions of those FCs are shown in Figs. 5 and 6, respectively. It is worth mentioning that both FCs have 6 rules and 12 rule conditions, however, their interpretability is not the same due to different fuzzy partitions.

In January 2007, an FC based on this project was installed into the bioaerosol detector and the preliminary tests proved its potential as an automatic reasoning mechanism.

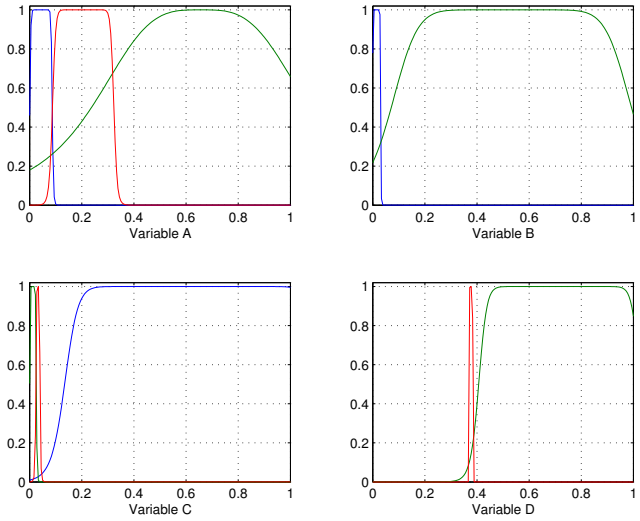

Fig. 6. Fuzzy partition with transparency penalty of 0.254. Relaxed covering occurs in variable A and there is a gap in variable D.

5. CONCLUSIONS

The goal of this work was to develop a model to be used as a reasoning mechanism in bioaerosol detector developed by Janka et al. (2007). It was desired that the model should have high true positive (TP) rate and low false positive (FP) rate. Furthermore, it was important for the sake of confidence in the model and of further development of the bioaerosol detector, that the developed model is as interpretable as possible. Therefore, the problem at hand was a multiobjective problem with conflicting objectives.

A hybrid genetic fuzzy system (GFS) was applied as an identification framework. It initialized the population with the help of crisp decision tree (DT) algorithm, which was clearly a better starting point for further optimization than commonly used random initialization. However, the initial fuzzy classifier (FC) was overly complex due to crispness of DT and due to the noise in the real-world data. Thus, merging of similar fuzzy sets took place and it led into some similar and inconsistent rules, which were heuristically removed. Then, the rest of the population was created, such that, the population was highly distributed in order to reduce the computational costs of multiobjective evolutionary algorithm (MOEA) optimization.

During MOEA optimization, the fitness of FCs was evaluated based on their FP and TP rates and transparency of their fuzzy partition. Even though the number of rules and rule conditions were not used as fitness objectives, the final population contained significantly less rules and rule conditions than the initial population. Furthermore, a vast amount of FCs with good tradeoff between the objectives were found. Because transparency of fuzzy partition was used as an objective, the obtained fuzzy partitions and rules were highly interpretable. In January 2007, an FC based on this project was installed into the bioaerosol detector and the preliminary tests proved its capability in distinguishing between safe and harmful bioaerosols.

REFERENCES

Janos Abonyi, Johannes A. Roubos, and Ferenc Szeifert. Data-driven generation of compact, accurate, and linguistically-sound fuzzy classifiers based on a decision-tree initialization. *International Journal of Approximate Reasoning*, 32(1):1–21, 2003.

Arie Ben-David. A lot of randomness is hiding in accuracy. *Engineering Applications of Artificial Intelligence*, 20 (7):875–885, October 2007.

O. Cordón, F. Gomide, F. Herrera, F. Hoffmann, and L. Magdalena. Ten years of genetic fuzzy systems: current framework and new trends. *Fuzzy Sets and Systems*, 141(1):5 – 31, January 2004.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

John F. Elder and Daryl Pregibon. A statistical perspective on knowledge discovery in databases. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 83–113. 1996.

Tom Fawcett. Using rule sets to maximize roc performance. In *IEEE International Conference on Data Mining*, pages 131–138, November 2001.

Antonio F. Gómez-Skarmeta, Fernando Jiménez, and Jesus Ibáñez. Pareto-optimality in fuzzy modeling. In *6th European Congress on Intelligent Techniques and Soft Computing EUFIT'98*, pages 694–700, Aachen, Germany, September 1998.

Peter P. Hairston, Jim Ho, and Frederick R. Quant. Design of an instrument for real-time detection of bioaerosols using simultaneous measurement of particle aerodynamic size and intrinsic fluorescence. *Jounal of Aerosol Science*, 28(3):471 – 482, April 1997.

Christian Haubelt, Jürgen Gamenik, and Jürgen Teich. Initial population construction for convergence improvement of moeas. In Carlos A. Coello Coello, Arturo Hernández Aguirre, and Eckart Zitzler, editors, *EMO 2005*, volume 3410 of *Lecture Notes in Computer Science*, pages 191–205, 2005.

Evan J. Hughes. Evolutionary many-objective optimisation: many once or one many? In *The 2005 IEEE Congress on Evolutionary Computation*, pages 222 – 227, September 2005.

Hisao Ishibuchi and Yusuke Nojima. Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning. *International Journal of Approximate Reasoning*, 44(1): 4–31, January 2007.

Hisao Ishibuchi, Tomoharu Nakashima, and Tadahiko Murata. Three-objective genetics-based machine learning for linguistic rule extraction. *Information Sciences*, 136 (1-4):109–133, 2001.

Hisao Ishibuchi, Yusuke Nojima, and Isao Kuwajima. Fuzzy data mining by heuristic rule extraction and multiobjective genetic rule selection. In *2006 IEEE International Conference on Fuzzy Systems*, pages 7824–7831, Vancouver, B.C., Canada, July 2006.

Kauko Janka, Riku Reinivaara, Juha Enroth, Jarmo Hytönen, Juha Tikkanen, Antti Rostedt, Matti Putkiranta, Marko Marjamäki, Jaakko Laaksonen, Jorma Keskinen, and Tarmo Humppi. Uv-fluorescence-detection based bioaerosol-warning concept with a background-aerosol sensor. In *The 9th Symposium on Protection against Chemical and Biological Warfare Agents*, Gothenburg, Sweden, May 2007.

Min-Seong Kim, Chang-Hyun Kim, and Ju-Jang Lee. Evolving compact and interpretable takagi-sugeno fuzzy models with a new encoding scheme. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 36(5):1006 – 1023, October 2006.

Silvia Poles, Yan Fu, and Enrico Rigoni. The effect of initial population sampling on the convergence of multi-objective genetic algorithms. In *MOPGP'06: 7th Int. Conf. on Multi-Objective Programming and Goal Programming*, Loire Valley (City of Tours), France, June 2006.

Foster Provost, Tom Fawcett, and Ron Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Fifteenth International Conference on Machine Learning (ICML 1998)*, pages 445–453, July 1998.

Pietari Pulkkinen and Hannu Koivisto. Identification of interpretable and accurate fuzzy classifiers and function estimators with hybrid methods. *Applied Soft Computing*, 7(2):520–533, March 2007a.

Pietari Pulkkinen and Hannu Koivisto. Fuzzy classifier identification using decision tree and multiobjective evolutionary algorithms. *International Journal of Approximate Reasoning (in press)*, 2007b.

Robin C. Purshouse and Peter J. Fleming. Evolutionary many-objective optimisation: An exploratory analysis. In *The 2003 Congress on Evolutionary Computation*, volume 3, pages 2066–2073, December 2003.

J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman Publishers, 2929 Campus Drive, Suite 260 San Mateo, CA 94403, 1993. ISBN 1-55860-238-0.

Hans Roubos and Magne Setnes. Compact and transparent fuzzy models and classifiers through iterative complexity reduction. *IEEE Transactions on Fuzzy Systems*, 9(4):516–522, 2001.

Magne Setnes and Hans Roubos. Ga-fuzzy modeling and classification: Complexity and performance. *IEEE Transactions on Fuzzy Systems*, 8(5):509–522, 2000.

Magne Setnes, Robert Babuŝka, Uzay Kaymak, and Hans R. van Nauta Lemke. Similarity measures in fuzzy rule base simplification. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 28(3):376 – 386, June 1998.

Christian Setzkorn and Ray C. Paton. On the use of multi-objective evolutionary algorithms for the induction of fuzzy classification rule systems. *BioSystems*, 81(2):101–112, 2005.

Vasanthi Sivaprakasam, Alan L. Huston, Cathy Scotto, and Jay D. Eversole. Multiple uv wavelength excitation and fluorescence of bioaerosols. *Optics Express*, 12(19): 4457–4466, 2004.

Hanli Wang, Sam Kwong, Yaochu Jin, Wei Wei, and K.F. Man. Multi-objective hierarchical genetic algorithm for interpretable fuzzy rule-based knowledge extraction. *Fuzzy Sets and Systems*, 149(1):149–186, January 2005.