The Effect of Antibody Morphology on Non-Self Detection

Johan Kaers[1], Richard Wheeler[2], and Herman Verrelst[1]

[1] Data4s Future Technologies,
Ambachtenlaan 13G, 3001 Heverlee, Belgium
{johan.kaers,herman.verrelst}@data4s.com
[2] Edinburgh Research and Innovation Ltd.
1-7 Roxburgh Street, Edinburgh EH8 9TA, Scotland
richard.wheeler@ed.ac.uk

Abstract. Anomaly detection algorithms inspired by the natural immune system often use the negative selection metaphor to implement non-self detection. Much research has gone into ways of generating good sets of non-self detectors or antibodies and these methods' time and space complexities. In this paper, the antibody morphology is defined as the collection of properties defining the shape, data-representation and data-ordering of an antibody. The effect these properties can have on self/non-self classification capabilities is investigated. First, a data-representation using fuzzy set theory is introduced. A comparison is made between the classification performance using fuzzy and m-ary data-representations using some benchmark machine learning data-sets from the UCI archive. The effects of an antigen data reordering mechanism based on Major Histocompatibility Complex (MHC) molecules is investigated. The population level effect this mechanism can have by reducing the number of holes in the antigen space is discussed and the importance of data order in the r-contiguous symbol match-rule is highlighted. Both are analysed quantitatively using some UCI data-sets.

1 Introduction

Anomaly detection algorithms based on the biological immune system have been applied to a variety of problems, including virus detection [7], network intrusion detection [12] [11] and hardware fault tolerance [3]. The T-cell maturation process is used as an inspiration for algorithms that produce a set of *change detectors* or *antibodies*. A *censoring* process removes a *detector* when it *matches* with a cell or *data-string* of the *self* from a large space of possible detectors. The remaining ones are then used to determine if an incoming data-string or *antigen* is part of the self or not. The time and space complexities of these algorithms have been extensively analysed and variations inspired on other immunological phenomena [14] and evolutionary computing [16] [10] have been proposed [12].

One common property is that they all use a binary or m-ary symbol string data-representation. In the biological immune system however, recognition between receptors and antigens is based on their three-dimensional shapes and

other physical properties [15]. Following the biological structure of the antibody-antigen binding process more closely could enable us to transpose the performance and adaptability of the natural immune system onto these computer algorithms. Therefore we investigate in this paper some modifications of the shape, data-representation and data-ordering of artificial antibodies. We will refer to these properties as the *antibody morphology*. They are analysed from the machine learning viewpoint and evaluated according to their usefulness as tools to enhance the classification capabilities.

The Artificial Life hypothesis is used because we assume that it is possible to model the biological immune system as a complex system of many interacting components. Similar to work in Artificial Chemistries [6] we abstract from natural molecular processes to investigate the fundamental dynamics of the complex system. An antibody morphology that uses Fuzzy Set Theory [20] is introduced. It captures the graded nature of the physical antibody/antigen match process and allows the non-self detection algorithms to handle data-sets with complex symbolic structures.

The paper is organized as follows. Section 2 looks at a data-representation inspired by fuzzy set theory and shows how the r-contiguous symbol matching rule can be modified accordingly. This fuzzy morphology is applied to some data-sets and statistically compared with the m-ary one. In section 3, the effect of using a data reordering method inspired by the *Major Histocompatibility Complex* (MHC) molecules is analysed. The importance of the data-order for the matching process is highlighted and the problem of *holes* in the non-self space is addressed by taking advantage of the population level effects resulting from using the MHC method. Section 4 contains the results and details of the various experiments using standard machine learning data-sets from the UCI repostiry [2].

1.1 Antibody Generation

Algorithms that generate antibodies make assumptions about their internal data-representation and therefore are tied to the antibody morphology. From the ones known in the literature [1], the *linear*, *greedy* and *binary template* algorithms build most heavily on the assumption of the binary (or m-ary [19]) string morphology. The ones based on generating random antibodies and/or genetic mutation operators (e.g. *exhaustive*, *NSMutation*) are more easily extended to arbitrary morphologies. The only requirement they have is that the morphology should allow for the generation of random antibodies and that a self to antibody matching scheme is present. Because of this independence of morphology, the *exhaustive* algorithm introduced by Forrest, Perelson et al. in [7] is used in the experiments included in section 4.

1.2 Machine Learning

Throughout the paper, non-self detection is considered as a 2-class classification problem, discriminating between self and non-self classes. The data-sets used

in the statistical analysis of the classifiers in section 4, are taken from the UCI Machine Learning archive [2]. These benchmark data-sets have well known properties and are widely used throughout the machine learning community, allowing comparisons with different types of classifiers [8].

2 Fuzzy Antibodies

The binding strength or *affinity* between an antigen and an antibody in the biological immune system depends on the physical and chemical forces acting at the binding site [15]. In the immune response, effective antibodies are the ones that bind tightly to an antigen and do not let go. In this section we capture the graded character of antibody-antigen affinity by modifying the bit-string morphology to include elements of fuzzy set theory. Other work also combines elements of artificial immune systems and fuzzy set theory. [13] shares our classifier point of view but uses another immune system model and another type of learning algorithm (AINE). [5] presents a hybrid function approximation algorithm that uses localized fuzzy models. [18] uses immune network theory and fuzzy soft-computing techniques for control and diagnosis of dynamical systems.

2.1 Data Fuzzification

Fuzzy set theory is used in various machine learning algorithms to cope with the inherent complexity often present in real-life datasets. It provides an intuitive means to represent data backed by a strong mathematical theory. As such, we use it here to extend the binary (and m-ary) string morphologies.

The *self set* S consists of a number N_s of independent data-strings of length l

$$s \in S = (x_1, \ldots, x_l)$$

where all x_i are values belonging to the *attribute* A_i. All attributes A_i have a number (n_i) of *fuzzy membership functions* or *FMFs* F_{ij} associated with them. The domain of these functions differs depending on the format and type of data the attribute is linked with. Every F_{ij} converts a value x_i to a *fuzzy membership value* between 0 and 1 that signifies a degree of membership to the function. In addition to this, a *match threshold* th_i is defined for every attribute.

The literature contains a large number of meaningful ways to define FMFs for various types of data. For example :

- *Continuous numbers*
  If the data is numeric and continuous from a given interval, the axis can be divided into a number of equal sub-intervals, each receiving one FMF. In this paper, we use triangle-shaped FMFs, centred in the middle of their sub-intervals. The amount of *overlap* between the triangles is determined by a parameter. A value of 1 means only its own sub-interval is covered, values > 1 and < 2 that the neighbouring intervals are partially covered, > 2 that further intervals are also covered. Figure 1 shows this kind of set-up for values from $[0, 100]$ and overlap $= 1.5$.

– *Categorical data*
  If the data is taken from a fixed set of unrelated symbols, every symbol gets an FMF that is 1 for that symbol and 0 for all the others. The FMFs reduce to identity functions, one for each symbol.
– *Textual data*
  Often, text data is structured in a way that allows meaningful fuzzy functions to be defined. e.g. an FMF acting on a date field can produce higher fuzzy membership values when the date is closer to a fixed moment in time. Spelling errors can be handled using an FMF based on digram matching [17].

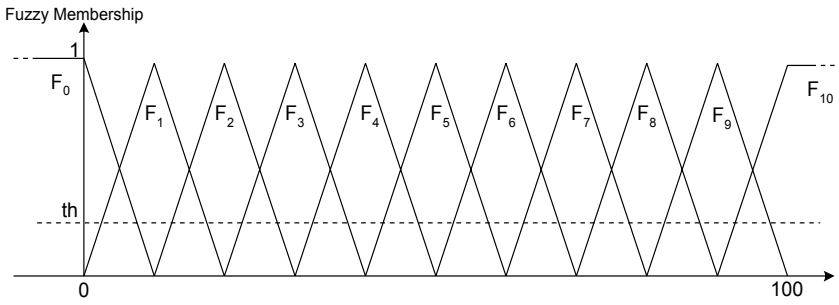

Fig. 1. The axis is divided up into overlapping domains by the FMFs. F_1 - F_9 cover the region between 0 and 100. F_0 and F_{10} capture everything outside this range. The dashed line is the fuzzy threshold for this attribute.

2.2 Fuzzy Matching

The *r-contiguous symbol* match rule that is used to determine if an antibody matches an antigen or self string is modified as follows. Assume the data contains l values. An antibody or detector d is a string of FMFs

$$d = (F_{1j_1}, \ldots, F_{lj_l}) \quad \forall i = 1 \ldots l : 1 \leq j_i \leq n_i$$

A string is said to *match* if there are r (the *match length*) contiguous positions where the data applied to the corresponding FMF exceeds it's fuzzy threshold. Call $a = (a_1, \ldots, a_l)$ the data-string.
a matches $d \Leftrightarrow$

$$\exists p, \forall q : p \leq q \leq p + r - 1 : F_{qj_q}(a_q) \geq th_q$$

Figure 2 illustrates this matching process with a data-string of length 5 and a match length of 3.

When all attributes are categorical with symbols 0 and 1 and the match threshold is set to 1, this set-up corresponds to the binary string morphology. The m-ary morphology is attained when all attributes are categorical with m categories, again with the match threshold set to 1. The thresholds define the specificity of a match. Low thresholds will result in a more approximate view on the data while higher ones constrain matching to more narrowly defined regions. The tuning of the FMF/Threshold combinations will therefore have a direct impact on the detection performance.

An immune response algorithm can take into account the *match affinity* f defined by

$$0 \leq f = \sum_{q=p}^{p+r-1} F_q j_q(a_q) - \sum_{q=p}^{p+r-1} th_q \leq r \tag{1}$$

that measures how strong the match between an antibody and an antigen is.

Section 4.2 gives some results obtained using this method together with all relevant details about the data-sets and algorithm parameters.

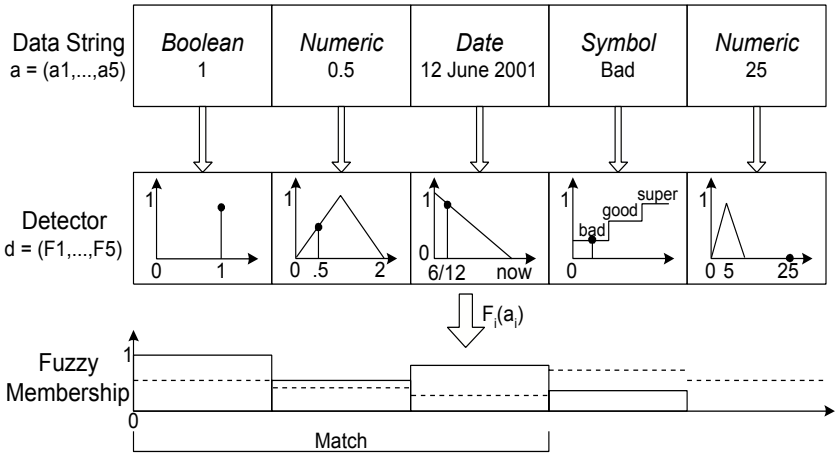

Fig. 2. Antigen - Fuzzy Antibody match with $l = 5$ and $r = 3$. The data is evaluated with the detector's FMFs (F_i) and compared to the fuzzy thresholds (dashed lines). The Fuzzy Membership Values at positions 1,2 and 3 exceed the thresholds. Therefore, there is a match.

3 Major Histocompatibility Complex

Major Histocompatibility Complex or *MHC* molecules play the role of antigen presentation agents in the biological immune system. They bind to parts of antigens and display these at the cell surface where they can be recognized by T-cells. There is a high degree of variability in the MHC molecules between

individuals, each mounting a slightly different immune response. This increases the population level immunity against diseases [15]. In this section, we look at a possible implementation of the MHC system and how it affects classification performance.

3.1 Population Level Effects

When using the r-contiguous match rule, there are always parts of the non-self space that are undetectable for any antibody. These *holes* in the non-self space occur where any detector for the hole will also match some self string and therefore is an undesirable one. This phenomenon was investigated in [4] and Hofmeyr [11] was the first to use the analogy with MHC molecules to combat this problem. By using a random permutation mask to produce a different data-representation in a number of local antibody sets, the holes also differ in every location. The likelihood that an antigen will fall in the cross-section of all local holes is lower than for the individual locations separately. Combining the local antibody sets can therefore result in a higher performance that cannot be attained with a single set. In section 4.3 this effect is illustrated using a synthetic data-set containing random bit-strings.

3.2 Attribute Order

Transposing the MHC system to the Artificial Immune System domain, it can be said to form a bridge between the data-representations of the antigen and the antibody. When using a string morphology (m-ary of fuzzy), the MHC molecules can be implemented as a permutation mask as first suggested in [11]. In this set-up, antigen data-strings are broken down into their components whose order is re-arranged according to the permutation mask before being presented to an antibody. Figure 3 illustrates this process.

When using the r-contiguous symbol match rule, the order of the attributes in the data-string is important. Per definition, the self/non-self patterns have to occur between attributes no further than r symbols apart. MHC permutations can therefore potentially destroy or enable the ability to capture the significant patterns in the data-set. When there are no dependencies between the attributes in the string, the re-arranging will not affect classification performance. In section 4.3 we illustrates this using some data-sets.

4 Results

All test were performed on a number of benchmark data-sets from the UCI machine learning database archive [2]. Continuous attributes were discretized for the m-ary symbol representation by dividing the interval of possible values into m equal sub-intervals and assigning a symbol to each one. For the fuzzy data-representation these same sub-intervals were used to space the overlapping Fuzzy Membership Functions as outlined in section 2.1

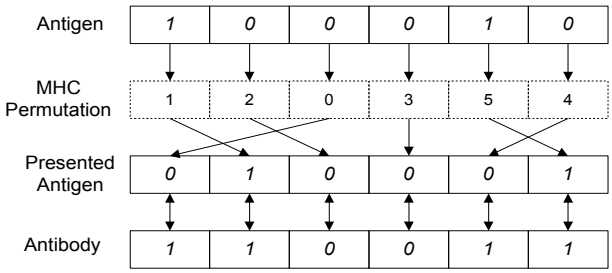

Fig. 3. Antigen presentation using a MHC permutation mask. The bits of the antigen are re-orderder according to the permutation mask before being matched to the antibody.

4.1 Data-Sets

The goal of the Iris data-set is to predict the type of iris plant using instances of 4 continuous attributes. This 3-class problem was converted to a 2-class one by grouping the "versicolor" and "virginica" types into 1 class. The Cancer and Hepatitis data-sets deal with medical diagnosis data. The Cancer data-set has 10 continuous attributes that are used to determine benign or malignant classes. The Hepatitis set uses 15 boolean and 5 continuous attributes to predict the survival of a patient. The Zoo Animals data-set classifies, using 15 boolean and 1 categorical attribute, a number of animals into 7 classes. We grouped together the non mammal classes, so the task became to distinguish between mammals and non mammals. The Mushroom set has 22 categorical values that define whether a mushroom is edible or poisonous. Finally, "Random Bit-String" is a data-set of 100 32-bit random bit-strings.

Data-Sets and Algorithm Parameters					
Data-set	**#Self**	**#Non-Self**	**#Antibodies**	**Length**	**Match Length**
Iris	50	100	300	4	2
Cancer	458	241	1000	9	3
Hepatitis	123	32	75000	20	4
Zoo Animals	41	59	100	16	7
Mushrooms	4208	3916	50000	21	8
Random Bit-String	100	-	200	32	8

Table 1. The data-sets and the size of the self- and antibody-sets derived from them. Length of the data-strings and the match-length used in r-contiguous matching

4.2 Fuzzy Morphology Results

Using the settings of table 1 we configured experiments to determine the classification error of Artificial Immune Systems using the m-ary and fuzzy morphologies. The results in table 2 were obtained from 10 runs of 10-fold cross validation. The number of continuous attributes is given, as well as the number of intervals in which they were discretized. We used the set-up of overlapping FMFs as outlined in section 2.1. The fuzzy threshold was set to 0.01 for these attributes and to 1.0 for the categorical attributes. Therefore the morphology reduces to the m-ary one when the fuzzy overlap is set to 1.0 (rows 1 and 3).

Classification Error					
Data-set	**Continuous**	**Intervals**	**Overlap**	**Error**	**Variance**
Iris	4	8	1.0	0.14899	0.09623
Iris	4	8	2.0	0.11744	0.03183
Cancer	9	10	1.0	0.08108	0.01449
Cancer	9	10	3.0	0.03567	0.00588
Hepatitis	4	5	1.0	0.20129	0.01257
Hepatitis	4	5	1.5	0.21415	0.02483
Zoo Animals	0	-	-	0.047	0.09597
Mushrooms	0	-	-	0.02093	0.00832

Table 2. Summary of classification errors on UCI repository data-sets using different antibody morphologies

The figures in table 2 seemed to indicate that fuzzy morphology can result in lower error-rates (row 2 $<$ row 1 and row 4 $<$ row 3). To confirm this, we performed paired t-tests of 30 runs of 10-fold cross-validation between a classifier using the m-ary (overlap 1.0) and one with the fuzzy morphology. Table 3 shows the average classification errors over the 30 runs for both morphologies and the outcome of the t-test. The *winner* column is the morphology with lowest average classifcation error and the *significance* column shows how certain we can be of this results according to the t-test. Given the results we can conclude that the data-sets that are dominated by continuous attributes (Iris and Cancer) could attain a lower classification error using the fuzzy morphology. The Hepatitis data-set did not show a significant difference in classification error when using the fuzzy morphology.

4.3 MHC Results

In order to illustrate the population level effects of MHC re-ordering we set up the following experiment using the Random Bit-String data-set as self set. One antibody set of 200 antibodies was generated without MHC reordering, while another one was split up in 4 sub-sets of 50 antibodies, each with a different

Morphology Comparison				
Data-set	**M-ary**	**Fuzzy**	**Winner**	**Significance**
Cancer	0.05912	0.03945	Fuzzy	98.14664%
Iris	0.14094	0.12081	Fuzzy	92.72657%
Hepatitis	0.20315	0.21496	(M-ary)	62.47845%

Table 3. Comparison between m-ary and fuzzy morphologies on data-sets.

random MHC permutation mask. All matching operations for these 4 sub-sets took into account the attribute re-ordering caused by their MHC permutation masks as illustrated in 3. For 1000 iterations bit-strings from the self set were picked at random and modified in 1 bit before being matched against both antibody sets. If a set failed to detect the change, the classification error was increased. This experiment was repeated 30 times and the classification errors from both sets used as input for a paired t-test. Table 4 shows that the statistical significance that the MHC set is better is very high. Since the data consists of random bit-strings, there is no significance in the order of the data-string, and the positive effect can be attributed to the population level effects discussed in section 3.1.

Other data-sets were investigated in a similar fashion. The Iris set was used to train an antibody set of 300 antibodies and another one containing 3 sub-sets of 100 antibodies with different MHC permutations. The Cancer set was tested with one set of 1000 versus 3 sub-sets of 333 with MHC. As in the previous section, 10-fold cross-validation was used and the classification errors of 30 such experiments were used in a paired t-test. The Iris set shows a significant difference favouring the classifier using MHC, while the Cancer set does not show a significant difference. We can conclude that for some sets the data-order and hole reducing effects of the MHC re-ordering are significant, whereas in other cases they show no positive or negative effect on the classification performance.

MHC Comparison				
Data-set	**No MHC**	**MHC**	**Winner**	**Significance**
Random Bit-String	0.3051	0.1548	MHC	99.99877%
Iris	0.1312	0.1083	MHC	99.93038%
Cancer	0.0335	0.0389	(No MHC)	56.31424%

Table 4. Comparison between classifiers without MHC re-ordering and with MHC re-ordering

5 Conclusions

In this paper we looked at properties related to the data-representation and data-ordering of antibodies, collectively called the antibody morphology. We introduced an alternative way of representing antibodies based on fuzzy set theory and showed how the r-contiguous symbol match rule can be adapted accordingly. Using the fuzzy morphology we trained a number of classifiers on data-sets from the UCI archive. Paired t-tests confirmed that using a properly tuned fuzzy morphology can have positive effect on the classification performance.

We discussed how a data-reordering mask inspired by MHC molecules can reduce the effect of holes in the antigen space and emphasized that the ordering of attributes has to take into account the specific properties of the data-set. Using data-sets of random bit-strings, the population level effect in the MHC reordering was shown to have a significant effect on the classification performance. We illustrated that depending on the inherent self/non-self patterns present in the data, the order in which the attributes occur can have an impact on the classification performance and therefore MHC reordering can influence the classification performance and robustness.

More research into methods to predict these effects and guide the MHC permutation process are needed to optimally benefit from them. Currently the authors are working on more formal ways to characterize the ideal data order inside the antibodies. Also, algorithms are being investigated that can inductively derive the optimal morphology and match parameters given an arbitrary data-set.

6 Acknowledgments

This work has been supported by Data4s Future Technologies, Leuven, Belgium. The authors would like to thank Prof. Peter Ross for the useful comments. Richard Wheeler would like to thank Marco Dorigo for his support and consideration.

References

1. M. Ayara, J. Timmis, R. de Lemos, L. de Castro, R. Duncan *Negative Selection: How to Generate Detectors*, in Proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS), 2002, p89-98.
2. C.L. Blake, C.J. Merz, *UCI Repository of machine learning databases*, http://www.ics.uci.edu/mlearn/MLRepository.html, Irvine, CA: University of California, Department of Information and Computer Science, 1998
3. R. O. Canham, A. M. Tyrrell, *A Multilayered Immune System for Hardware Fault Tolerance within an Embryonic Array*, in Proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS), 2002, p3-11.
4. P. D'haeseleer, S. Forrest, P. Helman. *An immunological approach to change detection : Algorithms, analysis and implications*, in Proceedings of the 1996 IEEE Symposium on Research in Security and Privacy, pages 110-119, IEEE Computer Society Press, Piscataway, New Jersey.

5. Y. Diao, M. Passino. Immunity-Based Hybrid Learning Methods for Structure and Parameter Adjustment, IEEE Transactions on Systems, Man and Machine, July 2000.
6. P. Dittrich, J. Ziegler and W. Banzhaf. *Artificial Chemistries - A Review*, Artificial Life VII, 3, 2001, p225-275.
7. S. Forrest, A. S. Perelson, L. Allen, R. Cherukuri, *Self-nonself discrimination in a computer.*, in Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy, Los Alamitos, CA : IEEE Computing Society Press, 1994.
8. Z. Frederick, *A Comprehensive Case Study: An Examination of Machine Learning and Connectionist Algorithms*, Masters Thesis, Brigham Young University, 1995
9. P. Helman, S. Forrest, *An efficient algorithm for generating random antibody strings*, Technical Report CS-94-07, The University of New Mexico, Albuquerque, NM, 1994.
10. R. R. Hightower, S. Forrest, A. S. Perelson, *The Evolution of Emergent Organization in Immune System Gene Libraries*, Proc. of the 6th Int. Conference on Genetic Algorithms, L. J. Eshelman (ed.), Morgan Kaufmann, 1995, p344-350.
11. S.A. Hofmeyr, *An Immunological Model of Distributed Detection and its Application to Computer Security*, Ph.D. Thesis, University of New Mexico, May 1999
12. J. Kim, P. Bentley, *An Artificial Immune Model for Network Intrusion Detection*, Proceedings of the 7th European Congress on Intelligent Techniques - Soft Computing (EUFIT'99). Aachen, Germany. September 13-19, 1999.
13. O. Nasraoui, F. González, and D. Dasgupta, *The Fuzzy Artificial Immune System: Motivations, Basic Concepts, and Application to Clustering and Web Profiling.* In IEEE International Conference on Fuzzy Systems, pages 711-716, Hawaii, HI, 13 May 2002. IEEE.
14. M. Oprea, S. Forrest, *Simulated Evolution of Antibody Gene Libraries under Pathogen Selection*, Proc. of the IEEE conference on Systems, Man and Cybernetics, 1998.
15. P. Parham, *The Immune System*, Garland Publishing/Elsevier Science, 2000
16. A. S. Perelson, R. Hightower, S. Forrest. *Evolution and Somatic Learning in V-Region Genes*, Research in Immunology 147, 1996, p202-208.
17. U. Pfeifer, T. Poersch and N. Fuhr. *Searching proper names in databases*, in Proceedings of HIM95: Hypertext-Information Retrieval and Multimedia, Konstanz, Germany, pages 259–275. Universitatsverlag Konstanz, 1995.
18. N. Sasaki, Y. Dote. *Diagnosis and Control for Multi-Agent Systems using Immune Networks*, Artificial Neural Networks in Engineering (ANNIE 2002), November 2002.
19. S. Singh, *Anomaly Detection Using Negative Selection Based on the r-contiguous Matching Rule*, Proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS), 2002, p99-106.
20. L. Zadeh, *Fuzzy sets*, Inf. Control 8, 338-353, 1965