Flexibility and Robustness of Hierarchical Fuzzy Signature Structures with Perturbed Input Data

B. Sumudu U. Mendis
Department of Computer Science
The Australian National University
Canberra, ACT 0200, Australia
sumudu.mendis@anu.edu.au

Tamás D. Gedeon
Department of Computer Science
The Australian National University
Canberra, ACT 0200, Australia
tom.gedeon@anu.edu.au

László T. Kóczy
Department of Telecommunication and Media Informatics, Budapest University of Technology and Economics, Hungary.
Institute of Information Technology and Elec. Eng.,
Széchenyi István University, Hungary
koczy@tmit.bme.hu

Abstract

We investigate the ability of fuzzy signature structure to cope with substantially reduced information as simulated by an experiment in which major branches were removed. We also investigated removal of significant number of individual data items. The results demonstrate that fuzzy signatures are robust under both of these conditions. This was our prediction, as fuzzy signatures are essentially hierarchical vector valued fuzzy sets designed for data sets and inconsistent substructures. To measure the succuss or otherwise of our experiments we introduced three new measures to compare results of two fuzzy signatures, being similarity, dissimilarity, and risk.

Keywords: Vector valued fuzzy sets, Fuzzy signatures, and Weighted aggregation.

1. Introduction

Soft computing research focuses mainly on identifying approximate models for decision support or classification where analytically unknown systems exist. Mostly, those systems consist of very complex structured, high dimensional data, and sometimes with interdependent features. Fuzzy logic approaches have become ideal for soft computing research because of the ability to assign linguistic labels [4] and to model uncertainty in most decision making and classification problems. But conventional fuzzy rule based systems suffer from high computational time complexity. Thus, their primary applicability still remains on real time control systems with few dimensions of input variables and simply structured data.

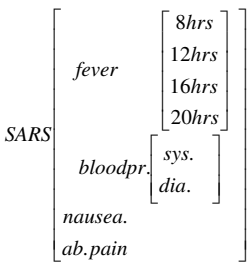


Fig. 1 Example of a Simple Fuzzy Signature

The hierarchical fuzzy signatures structure is a novel concept that can be used to find the degree of similarity or dissimilarity of objects which contain complex structured data (fig.1), for classification or decision making. Fig.1 shows a simple example of a fuzzy signature structure, which contains a complex structured data [2] & [3]. In *Kóczy et a*l [6] the fuzzy signature concept has been introduced and combined signatures with vector valued fuzzy sets [5] to develop the theoretical aspects of fuzzy signatures. The term fuzzy signature is unfortunately also used in some publications in network attack detection in the context of security using fuzzy association rules [7], but the two concepts are not related to each other. Also, the importance of the concept of fuzzy signatures in

data mining has been discussed in *Vámos et al* [8]. In *Wong et al* [2] & [3] the construction of hierarchical fuzzy signature structure from data has been discussed. In our early work [1], the aggregation of fuzzy signatures, special benefits of fuzzy signatures for decision making and classification, and some differences between hierarchical fuzzy signatures and sparse hierarchical fuzzy rule based systems were discussed. Moreover, a new inference method for fuzzy signatures called weighted aggregation and a simple new aggregation function called MA (Maximum_Average) were introduced.

The fuzzy signatures concept has been developed to describe the objects in somewhat the same way humans model the problems. That is, their hierarchically structured fuzzy sets contain the interconnectedness of features of the object by their hierarchical structure, in fuzzy signatures. Also, the interconnected relationship between higher and lower levels of the fuzzy signature structure is derived by a set of qualitative measures, which are not necessarily homogeneous [1] & [6]. In addition to that, humans have an ability to make decisions when some components of data were removed or missing from input data. The Fuzzy signatures concept has been developed with this objective in mind [2]. Thus, fuzzy signatures are capable of handling systems with complex structured data and sometimes with input data missing or removed. In section 2, we briefly discuss the theoretical background of fuzzy signatures and weighted aggregation method, which are used in two experiments.

In some situations it is necessary to reduce or aggregate information to become compatible with information obtain from another source, where some detail variables are missing or simply been removed from the input data. In such situations human experts still can make decisions based on the existing knowledge. For example, medical practitioners can make some decisions based on the available data for a patient in urgent situations. As explained in the previous paragraph, the fuzzy signatures concept approach to problem solving is similar to the way humans do. Thus, the fuzzy signature can be used to make decisions when data is missing or removed from the inputs.

In such situations there are two methods of approximating the results based on existing input data. First method is to use an optimal fuzzy signature structure, which is suited to the existing data. This optimal fuzzy signature structure can be derived (by which we mean modifying the existing signature structure to better suit the particular data point; currently we do this manually as we do not yet have an automated technique) from the maximal common fuzzy signature structure. In such situations, it has been pointed out in [1] that importance of finding the limits and constraints of deriving an optimal fuzzy signature structure from the maximal common fuzzy signature structure in order to avoid approximating inaccurate results from existing data. Therefore, in experiment 1, we investigated the accuracy of the results of derived fuzzy signature structures versus the results of the maximal common fuzzy signature structure for a certain set of test data. In order to compare these results, three measures have been proposed to find the similarity, dissimilarity and risk of results of two fuzzy signature structures, which contain the same skeleton. Further these three measures can be used to find the flexibility of the maximal common fuzzy signature structure for a particular situation where data is missing or removed from the inputs.

On the other hand, as the second method, instead of using an optimal fuzzy signature structure, the same fuzzy signature structure still can be used by filling the missing component of the original fuzzy signature using the existing input data. Thus, our proposed measures similarity, dissimilarity, and risk can be used to find the robustness of the fuzzy signatures in such situations.

The results of our experiment 2 show that robustness of the fuzzy signature is very high when the missing or removed data is properly filled. Also, during the experiment some filling mechanisms for missing components of the fuzzy signature have been evaluated.

For all the experiments, described in each section, example problems have been undertaken from medical diagnosis. Also, when aggregating fuzzy signatures for a final result the weighted aggregation method discussed in [1] has been used.

Our goal is to investigate the ability of fuzzy signatures to deal with problems, which contain complex structured input data, such as medical and economic diagnosis.

2. The Fuzzy Signatures and Weighted Aggregation Method.

In this section we discuss the theoretical aspects and some examples of vector valued fuzzy sets, fuzzy signatures and weighted aggregation method.

2.1. Vector Valued Fuzzy Sets.

A vector valued fuzzy sets is a more generalised form of I- fuzzy sets (interactive fuzzy sets) [5]. Simply, a vector valued fuzzy set, $\bar{A}$ =(x, $q_{\bar{A}}$) can be defined as, $q_{\hat{A}} : X \rightarrow [0,1]^n$, where $n \in N$. For example, the following membership degree matrix, $M_{\bar{A}}$ (Fig.2) represents patient A's

$M_{\bar{A}}$ =

	Slight	Mod	High
8 Hrs fever	0.8	0.7	0.1
12 Hrs fever	0.2	0.7	0.1
16 Hrs fever	0.1	0.2	0.8
20 Hrs fever	0	0.1	1

Fig.2. Membership degree matrix

temperature count at 4 times (*8hrs, 12hrs, 16hrs, and 20hrs*) of a day. Each temperature count is represented by doctor's diagnosis levels (*slight, moderate, high*). The membership degree $q_{\overline{12Hrs.}}(Mod)$ (fig.2.), means patient A's fever at 12hrs has 0.7 grade to doctor's *moderate* fever diagnosis level.

2.2. Fuzzy Signatures

Fuzzy signatures are vector valued fuzzy sets, where each vector component can be a further vector valued fuzzy set [6] & [8]. A fuzzy signature, *S* can be defined as,

$$s : X \rightarrow [a_i]_{i=1}^{k}, \quad where \quad a_i = \begin{cases} [0,1]; & if \quad leafe \\ [a_{ij}]_{j=1}^{k_i}; & if \quad branch \end{cases}$$

The Fig.3 shows, an example for a fuzzy signature [1], which represents patients SARS condition (degree of having SARS). The patients SARS condition is given by four symptoms namely

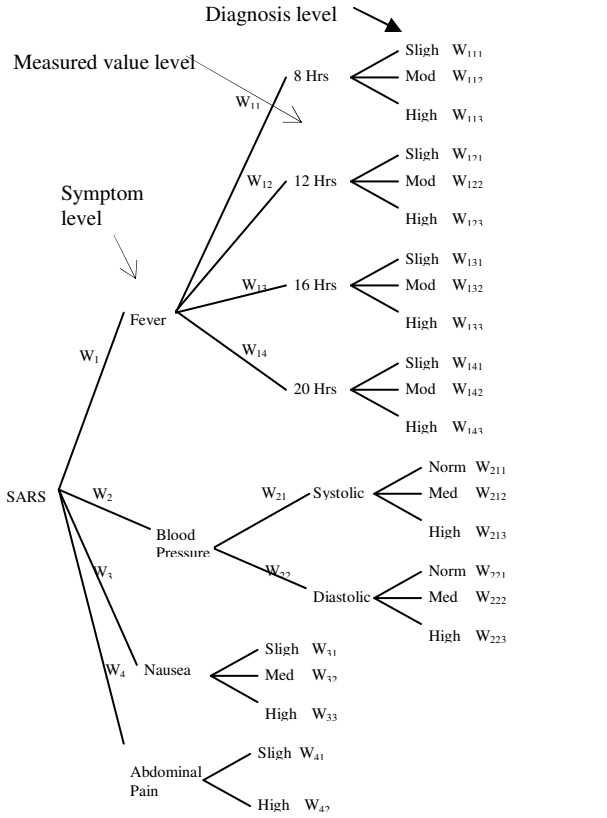


Fig.3. SARS Fuzzy Signature Structure

fever, blood pressure, nausea, and *abdominal pain*. These symptoms further represented by some measured values. For an example *fever* is given by four tests at different times, *8hrs*, *12hrs*, *16hrs*, and *20 hrs*, of the day. Also, further these measured values are represented by physician's diagnosis levels. For an example, *8hrs* fever count is given by three diagnosis levels, namely *slight fever, moderate fever,* and *high fever*, according to physician.

2.3. Comparison of two Fuzzy Signatures

Two fuzzy signatures can be compared, even with different structures but with same common skeleton, to find the similarity or dissimilarity of objects. This is one of basic advantage of fuzzy signatures, which is not possible in conventional rule base fuzzy systems. The fuzzy signatures with a common skeleton means that they must be sub sets, according to their structure, of a maximal common fuzzy signature structure. In order to compare two fuzzy signatures we use the following three equations given in [6],

$$S = (S_1 \wedge S_2) \vee (\overline{S}_1 \wedge \overline{S}_2) \quad (1)$$

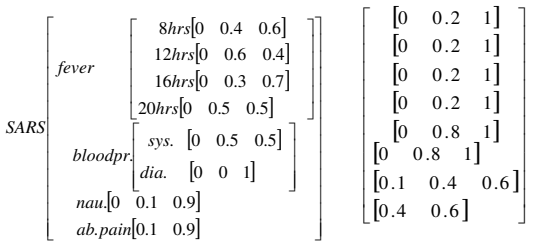
$$S = (S_1 \wedge S_2) \quad (2)$$

where S, S_1, and S_2 are fuzzy signatures and $a \wedge b = ab$, $a \vee b = a + b - ab$, and $\bar{a} = 1-a$, and

$$dist(v_1, v_2) = \|v_1 - v_2\| \quad (3)$$

where v_i are vector valued fuzzy sets and they must be in the same skeleton structure and the $\| \ \|$ means some norm operator.

2.4. Weighted Aggregation Method.



Inputs for SARS fuzzy Signature. Weights

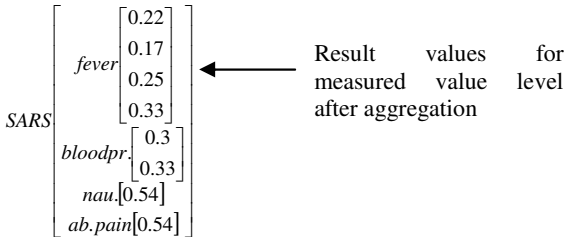


Fig.4. Weighted aggregation method

The weighted aggregation method is introduced to reflect the fact that some branches may contribute more to the final result than the others in the same level [1]. As an example, contribution of *slight fever*, *moderate fever*, and *high fever* (Fig.3) to the final SARS condition can be expressed linguistically as *"less", "somewhat"*, and *"more"*. Therefore, the weights w_{111}, w_{112}, and w_{113} in Fig.3 have been configured according to these linguistic expressions. The weights in the diagnosis level (Fig.3) represent the degree of relevance of diagnosis to the measured value level. Similarly, weights in measured value level represent the degree of relevance of measured values to the symptom level. Finally, the weights in the symptom level represent the degree of relevance of symptom to the SARS condition. Thus, the weighted aggregation method employs extra knowledge of the relevance of the lower levels to the higher level of the fuzzy signature structure.

The above example (fig.4) shows how weighted aggregation can be used to aggregate one level of a fuzzy signature. The SARS Fuzzy signature structure and aggregation functions AVG for fever and blood pressure and MAX for nausea and abdominal pain has been used to aggregate the diagnosis level in this example.

3. Experiment 1: The Accuracy of the Optimal Fuzzy Signature Structure.

As discussed earlier, the first method of approximating the results based on existing input data is to derive, an optimal fuzzy signature structure better suite to the particular data point, from the maximal common fuzzy signature structure. This is a more permanent solution, especially when some branches of data were removed from inputs. In such situations, that the importance of finding the limit of reducing the maximal common fuzzy signature structure and the accuracy of the derived fuzzy signature structure, has been pointed out in [1]. This is important to avoid predicting the inaccurate results. Therefore, the first experiment has been set up to investigate how far to reduce the maximal common signature structure.

The SARS fuzzy signature in fig.3 has been taken as the example maximal common fuzzy signature structure. Two derived fuzzy signatures have been formed to represents two different situations, where some branches were removed from input data. The abdominal pain branch has been removed from fig.3 to form the first derived fuzzy signature structure (changed signature 1). Now, the abnormal pain branch no longer exists in the new signature structure (fig.5).

Next, the systolic blood pressure branch has been removed from the original signature (fig.3) to form the second derived fuzzy signature structure (changed signature 2). Now, the blood pressure branch no longer exists and diastolic pleasure is only a branch in level 1 of the new signature structure (fig.6). This situation similar to a

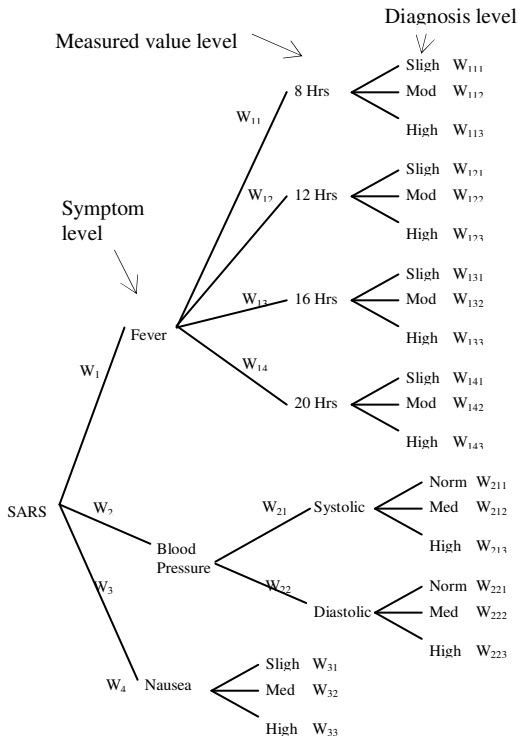


Fig.5. Changed SARS Fuzzy Signature Structure 1

situation, were the systolic blood pressure part is removed from the input data.

We used 4000 records of test data, which were derived from real details using some randomized components. These 4000 data contain SARS, normal, pneumonia, and hypertonia patient data. Each condition contains 1000 records of data. The best aggregation function (MAX, MAX, MIN), which was found in [1], for the SARS fuzzy signature is used for the experiments.

As the next step of the experiment, two methods of finding similarity and dissimilarity between results of two fuzzy signatures have been considered. Similar to the distance measure (3) of vector valued fuzzy sets, we can define the dissimilarity between results of two fuzzy signatures as follows:

$$Dsim(S_1, S_2) = |S_1 - S_2| \quad (4)$$

where S_1, and S_2 are aggregated results of two fuzzy signatures and L_1 norm has been used for simplicity. Also, as explained in [6], the dissimilarity or distance between two fuzzy signatures is very small if the similarity degree between them is very high. Thus, similarity can be defined as,

$$Sim(S_1, S_2) = 1 - Dsim(S_1, S_2) \quad (5)$$

The above measures can not be used in this form when a set of data is in hand for training or evaluation. Therefore, these simple measures of similarity and dissimilarity have been further extended into finding the average distribution of similarity and dissimilarity, called **average similarity** and **average dissimilarity**, between the results of two fuzzy signatures when set of test data is in hand for evaluation. Assume that set of data T with *n* records are given, now the average similarity can be defined as,

$$Sim^T(S_1, S_2) = \frac{\sum_{i=1}^{n} Sim(S_1^i, S_2^i)}{n} \quad (6)$$

where S_1^i means result of the fuzzy signature S_1 for the i^{th} data record. Average dissimilarity is defined as,

$$Dsim^T(S_1, S_2) = \frac{\sum_{i=1}^{n} f(i)}{n} \quad (7)$$

$$\text{where } f(i) = \begin{cases} 1, & \text{if } Dsim(S_1^i, S_2^i) \geq m \quad \text{where} \quad 0.5 \leq m \leq 1 \\ 0, & \text{otherwize} \end{cases}$$

As given by the equations (6) and (7), the average similarity and average dissimilarly are not complements of each other. Hence, they give two different important views of the results for comparisons. The average similarity gives an idea of how a derived fuzzy signature's best results are averagely similar to that of the initial signature results. During the experiments it has been observed that when average similarity of the derived signature is high compared to the maximal common fuzzy signature but still it gives some considerable amount of incorrect results. Therefore, the idea behind the average dissimilarity measure is to locate results which are incorrect but still on average are similar, using accuracy level *m*, of the derived fuzzy signature over initial fuzzy signature results. Thus, it gives the average presence of inaccurate results according to level *m*. The parameter *m* defines the margin of the accurate results.

The fig.7 shows the results of the experiment 1. During experiment 1 we compared the similarity

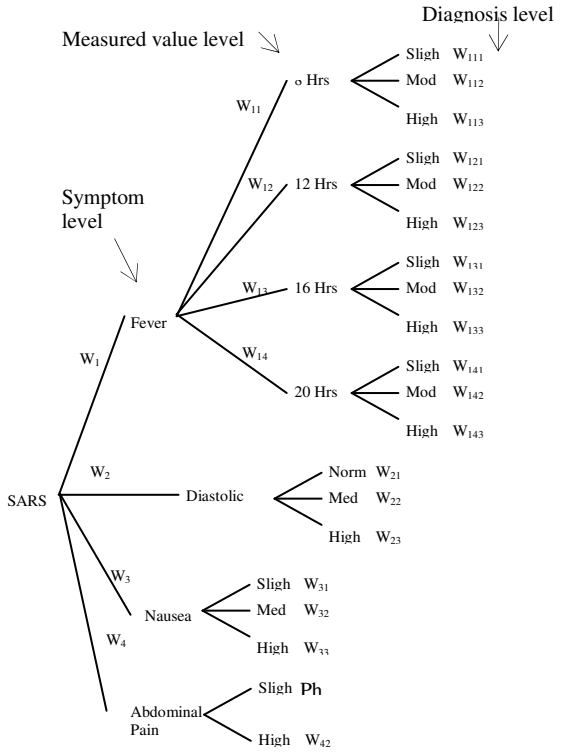
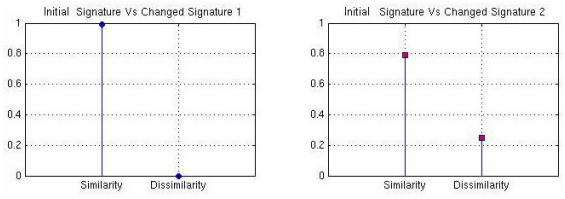


Fig.6. Changed SARS Fuzzy Signature Structure 2

and dissimilarity of the results of the 2 derived fuzzy signatures over the results of the maximal common fuzzy signature (fig.3).

As graph 1 in fig.7 shows, the similarity of the results of the derived signature 1 is very high (almost 1) and dissimilarity of that of derived signature is very low (almost 0) compared to the original signature. Thus, it can be concluded that the risk of using the derived signature 1 is negligible. But, according to the graph 2 in fig.7 the similarity of the results of the derived signature 2 is high (0.8) and also average dissimilarity is fairly low (0.25) and its significant is unclear. In addition to that, it has been observed that derived fuzzy signature 2 gives incorrect results for most of high blood pressure patient data. Therefore, the risk of using the derived fuzzy signature 2 is high, especially for medical diagnosis. Therefore it has been observed that these two concepts, ie. similarity and dissimilarity are not enough to identify the risk of the exact situation. Also, it is intuitive that the acceptable amount of the risk of the derived fuzzy signature depends on the practical application.



Graph 1 Graph 2

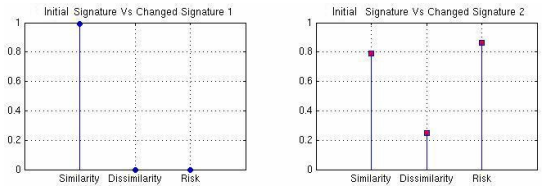
Fig.7. Similarity and Dissimilarity of changed fuzzy signatures

Therefore, the need of an additional measure which models and shows significant risk of the derived fuzzy signature structure is raised. In addition, as the secondary objective, this function should be capable of showing the risk depending on the application area. Therefore, a new measure called risk of the fuzzy signature, given by equation (8), has been proposed to achieve the above objective. The new function shows the risk of the results of one fuzzy signature over the other fuzzy signature results. Now, the degree of risk of the result of the derived signature over the original signature can be defined as:

$$\text{Risk}(S_1, S_2) = \left| 1 \quad - \quad \frac{1}{e^{\lambda \ \text{Dsim}^{T}(S_1, S_2)}} \right| \tag{8}$$

were $\lambda \geq 1$ and S_1 and S_2 two fuzzy signatures, $\text{Dsim}^{T}(S_1, S_2)$ is the average dissimilarity between derived and initial fuzzy signatures for data set T. The parameter λ is the most important parameter in the risk function. It has been setup to adopt the behaviour of the risk function according to the application. The above risk curve with λ =8 has been used for all remaining experiments on the SARS fuzzy signatures in this paper, by having an understanding that medical diagnosis accepts very small risk.

Fig.8 shows the results of the same experiment with additional risk function. Now, using risk it can be clearly describe the exact situation. Now, the graph 1 in fig.8 shows the risk of using the changed signature 1 is very low, but that of the changed fuzzy signature 2 is very high even with the high similarity and low dissimilarity. Now, according to fig.8, the derived fuzzy signature structure 1 can successfully be used to approximate the accurate results. And the changed signature structure 2 can not be used to approximate accurate results.



Graph 1 Graph 2

Fig.8. Similarity, dissimilarity, and risk

In the final test of the 1st experiment we formed 14 derived fuzzy signature structures by assuming combinations of symptoms were removed from input SARS data. The combinations are shown in the table 1. For an example to form the first derived fuzzy signature the abdominal pain branch has been removed from initial signature and to form the 3rd derived fuzzy signature both abdominal pain and nausea branches have been removed.

Table 1. Combinations of symptoms.

Derived Signature	Abdom. Pain	Nausea	High	Fever
1	0	1	1	1
2	1	0	1	1
3	0	0	1	1
4	1	1	0	1
5	0	1	0	1
6	1	0	0	1
7	0	0	0	1
8	1	1	1	0
9	0	1	1	0
10	1	0	1	0
11	0	0	1	0
12	1	1	0	0
13	0	1	0	0
14	1	0	0	0

The idea behind this experiment is to find out the accuracy of the derived SARS fuzzy signature structures when major components were removed from the inputs. Fig.9 shows similarity, dissimilarity, and risk of the results of these 14 derived signature structures compared to the initial SARS fuzzy signature structure.

According to Fig.9 it can be observed that risk of the derived fuzzy signature is very high when both nausea and abdominal pain branches are removed. It can be concluded that most of the results of the SARS fuzzy signature are highly depend on symptom nausea and abdominal pain branches. Thus, these 3 measures are useful to identify the behaviour of the fuzzy signature structures in the training phase.

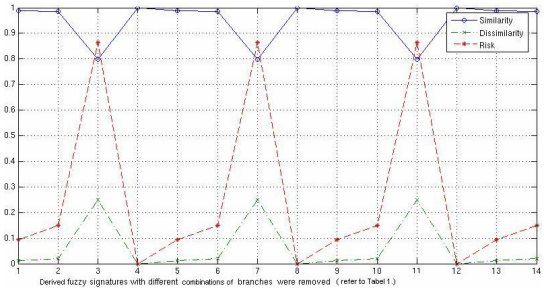


Fig.9. Similarity, dissimilarity, and risk of 14 combinations of derived fuzzy signature structures.

4. Experiment 2: The Robustness of the Original Fuzzy Signature Structure.

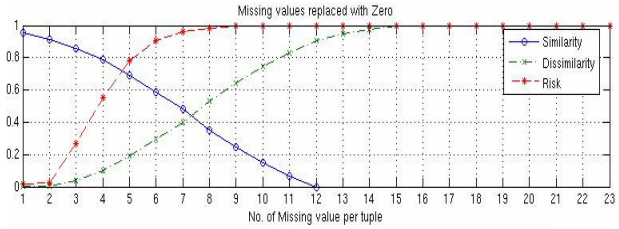


Fig. 10. Missing values replaced with zero

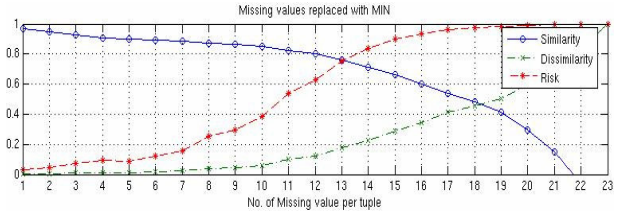


Fig. 11. Missing values replaced with Min

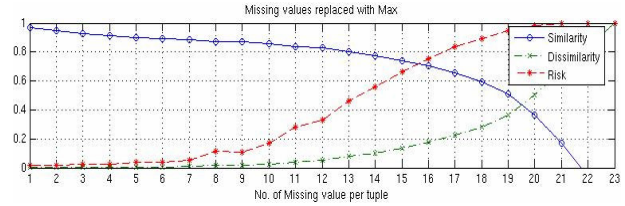


Fig. 12. Missing values replaced with Max

As described in the introduction, instead of using an optimal fuzzy signature, the maximal common fuzzy signature structure can still be used by filling in the missing components of the original fuzzy signature using the existing input data. In the literature, in Wong [3] the average of the branch has been used to fill in the missing values of the fuzzy signatures. In our second experiment four additional

filling mechanisms have been used for evaluation. That is filling the missing components of the fuzzy signature using zero, one, minimum of the branch, and maximum of the branch.

The scope of the second experiment is to find the robustness of the fuzzy signature structure when missing values of the input data were filled using above five methods. Our new measures, similarity, dissimilarity, and risk, have been used to find the robustness of the fuzzy signature in such situation.

The same set of test data, which is used for the experiment 1, has been used for the experiment 2. In order to generate missing values input data sets from original data set, randomly selected n ($0<n<24$) fields were removed from the each record, which contain 24 fields, of the original input data set. In this way 23 different input data sets were created and they have been numbered from 1 to 23. For an example, to create the data set two, 2 (n=2) randomly selected components were removed from each record of 4000 records. Now, each record of derived data set two contains 2 missing values. Similarly, each record of derived data set 10 contains 10 missing values.

The graphs in fig.10 -12 show only the 3 good filling mechanisms (zero, min, and max) of the experiment 2. All graphs show the similarity, dissimilarity and risk of the results of all 23 data sets compared to the results of the original data set.

Overall, as the figures show, the robustness of the fuzzy signature is very high when missing components of the fuzzy signature are filled with maximum value of the particular branch. This can be explained as the robustness of the fuzzy signature can be kept high when the missing components are filled in using an appropriate mechanism.

5. Conclusion.

One of the most useful benefits of fuzzy signatures over conventional rule base fuzzy systems is to deal with removed parts or missing individual input data points has been discussed. First, three measures for finding similarity, dissimilarity, and risk of the result of fuzzy signatures in such situations were proposed. The first experiment examines and reported the flexibility of the fuzzy signature structure to available data. Also, it has been discussed that these three measures can be used to find the behaviour of the fuzzy signature structure in the training phase. The second experiment concluded that the robustness of the fuzzy signature structure is very high when the missing individual data points have been filled properly.

Reference

[1] BSU Mendis, TD Gedeon, L.T. Kóczy, "Investigation of Aggregation in Fuzzy Signatures", *3rd Int. Conf. on Comp. Intelligence, Robotics and Autonomous Systems*, Singapore, 2005.

[2] K.W. Wong, A. Chong, T.D. Gedeon, L.T. Kóczy, T. Vámos, " Hierarchical Fuzzy Signature Structure for Complex Structured Data," *Proc. of Int. Symp. on Comp. Intelligence and Intelligent Informatics 2003 (ISCIII'03)*, 2003, Tunisia, pp 105-109.

[3] K.W. Wong, T.D. Gedeon, L.T. Kóczy, "Construction of Fuzzy Signature from data," *Proc. of IEEE Int. Conf. on Fuzzy Systems*, 2004, vol. 3, pp 1649-1654.

[4] L.A. Zadeh, "Fuzzy Algorithm", *Information and Control*, vol. 12, 1968, pp. 94-102.

[5] L.T. Kóczy, "Vector Valued Fuzzy Set," *Busefal, ete* 1980, pp. 41-57.

[6] L.T. Kóczy, T. Vámos, G. Biró, "Fuzzy Signatures," *Proc. of EUROFUSE-SIC '99*, 1999, pp. 210-217.

[7] M Manic, B Wilamowski, "Fuzzy Preference Approach for Computer Network Attack Detection," Proc. *of Int. Joint Conf. of Neural Networks*, 2001.

[8] T. Vámos, L.T. Kóczy, G. Biró, "Fuzzy Signatures in Data Mining," *Proc. of the joint 9th IFSA World Congress*, 2001, pp. 2842-2846.