Finding input sub-spaces for Polymorphic Fuzzy Signatures

A. H. Hadad, T. D. Gedeon, B. S. U. Mendis, *School of Computer Science, Australian National University*

Abstract— **A significant feature of fuzzy signatures is its applicability for complex and sparse data. To create Polymorphic Fuzzy Signatures (PFS) for sparse data, sparse input sub-spaces (ISSs) should be considered. Finding the optimal ISSs manually is not a simple task as it is time consuming; moreover, some knowledge about the dataset is necessary. Fuzzy C-Means (FCM) clustering employed with a trapezoidal approximation method is needed to find ISSs automatically. Furthermore, dealing with sparse data, we should be mindful about choosing a reliable trapezoidal approximation method. This facilitates the optimal ISS creation for the data. In our experiment, two trapezoidal approximation methods were used to find optimal ISSs. The results demonstrate that our version of trapezoidal approximation for creating ISSs result in an PFS with lower mean square error compared to the original trapezoidal approximation method.**

Index Terms— **Fuzzy Signatures, Polymorphic Fuzzy Signatures, WRAO, Fuzzy C-Means, Input subspace clustering, Trapezoidal Approximation**

I. INTRODUCTION

In the real world, we often receive data from sources that are sparse and hierarchical. Polymorphic Fuzzy Signatures (PFS) is a set of hierarchically structured recursive vectors. Each vector component may represents the next level of the hierarchical structure or it can be a leaf node [13]. PFS is inherently sparse [13] and the concept derived from the idea Fuzzy Signatures, which is developed to model complexly structured data [11]. It employees a set of hierarchically organized aggregation functions to find the final degree of match [22], [23], [13]. As an advantage, PFS is applicable for comparing the degree of similarity or dissimilarity of two slightly different objects. However, these two objects should have the same PFS skeleton. Additionally, PFS is capable of dealing with missing input data [22], [23], [16]. *Mendis et. al.* introduced generalized Weighted Relevance Aggregation Operator (WRAO) for PFS [17]. WRAO allows for the learning of both aggregation function and weighted relevance at the same time for one node in the PFS structure. Therefore, WRAO simplifies the learning of PFS models from data. *Mendis et. al.* demonstrated a successful way to extract WRAO for PFS [17], [18] based on the Levenberg-Marquardt (LM) optimization method [12]. Experiments presented by *Mendis et. al.* confirmed that the LM method can learn both

aggregations and weighted relevancies PFS [18]. In all the studies related to PFS that are mentioned here, the process of creating Input Sub-Spaces (ISSs) was performed manually. PFS has been successfully applied to decision making and classification problems Computational Intelligence [15], [14], [26].

In this paper, instead of generating the ISSs manually [22], [23], [17], [18], [19], we utilized a novel automatic approach to find fuzzy ISSs for PFS. This is a significant improvement in generating PFS, because it facilitates the creation of PFS for datasets with an unknown nature. This approach is based on fuzzy c-means (FCM) clustering [3] and a modified version of trapezoidal approximation method [21]. In this approach, we have introduce a novel version of Tikk's trapezoidal approximation method for sparse data [21]. We applied Tikk's and our trapezoidal approximation methods for the high salary selection problem [5]. The high salary selection Problem was discussed in [5]. In this problem the aim is to find the degree of relevance of having a high salary with *contacts*, *age*, and *work experience* of an employee. Finally, we have compared the trapezoidal approximation methods for this problem.

II. POLYMORPHIC FUZZY SIGNATURES

PFS can represent, compare and classify objects with complex hierarchical structures and interdependent features. Its hierarchical organization represents the structural complexity of a problem. PFS can be employed to approximate the global preference relation of a decision problem.

A. Vector Valued Fuzzy Sets

The PFS concept is a generalization of the Vector Valued Fuzzy Sets (VVFS) concept. The early work of Kóczy introduced the Vector Valued Fuzzy Sets concept [10] . The VVFS is a distinctive form of an *L*-fuzzy set, and can be denoted in the following form:

$$\mu_A : X \to [0,1]^n . \tag{1}$$

Where A is a fuzzy set and $L = [0,1]^n$ is in (1) and thus VVFS is *L*-fuzzy. The qualitative meaning of an object is represented by the quantities of the VVFS. The notation of the vector valued fuzzy set A is written as $\overline{A} = (x, \mu_A)$ and the membership function μ_A can be defined as, $\mu_A : X \to [0,1]^n$, where $x \in X$.

B. Polymorphic Fuzzy Signatures Structure

The fuzzy signatures concept provides a solution for the rule explosion problem in fuzzy logic because it is hierarchically structured and sparse. Fuzzy signatures definition is as follows [11], [13]:

Definition 1: Fuzzy signatures is a VVFS, where each vector component is another VVFS (branch) or an atomic value (leaf), and denoted by,

$$\mu_A : X \to [a_i]_{i=1}^k \left(\equiv \prod_{i=1}^k a_i \right). \qquad (2)$$

$$where \; a_i = \begin{cases} [a_{ij}]_{j=1}^{k_i} & ; if \; branch \\ [0,1] & ; if \; leaf \end{cases}$$

and Π describes the Cartesian product.

In the concept of Fuzzy Signatures, each incoming data point will become a Fuzzy Signature. PFS is an optimized approach to the initial idea of fuzzy signatures and one PFS can represents a set of individual data points. This has been achieved by replacing atomic leaf nodes in fuzzy signature structure (see definition 1) by fuzzy sets and using specialized aggregation functions such as WRAO [13]. Therefore, PFS concept is capable of solving complex problems in a effective way. In this paper, we discussed the concept of PFS as a practical approach that organizes and aggregates data hierarchically. Because of the hierarchal structure of PFS it is computationally more efficient.

III. Automatic ISSs finding

As mentioned before, we employed FCM for automatic ISSs finding. There are two main issues that need to be considered when creating the ISSs. One issue is finding the optimal number of sub-spaces in ISSs for each input, and other is finding the boundaries between each individual sub-space of ISSs.

A. Optimal number of sub-sapces

For each dataset, there are a number of inputs. Each input can be divided into several sub-spaces. The set of sub-spaces that are created by analyzing an individual input's data, are that input's sub-spaces which is called input sub-spaces (ISSs) in this text. For example in HSS problem, we can have three age groups that are the age sub-spaces. In our approach we employed FCM to define individual sub-spaces for each input; therefore, each individual sub-space (ISS) is equivalent to a fuzzy-cluster and each ISSs is equivalent to a set of clusters for an input. One of the issues with defining ISSs, is the optimal number of sub-spaces. Considering the equivalency of sub-spaces and clusters in our approach this can be addressed as a cluster validity problem. There are a variety of surveys on cluster validity in the literature [9], [8], [7], [27], [6]. *Halkidi et. al.* introduced two groups of cluster validity measures [6]. External cluster validity measures focus on the labels of the data in the dataset and validates the clusters. These measures of useful for the cases where there exists knowledge about the data [7]. Internal cluster validity measures focus on internal statistical features of the clusters for their validation. In this study, we employed two of traditional measures from the internal measures for FCM that demonstrated by *Sugeno et. al and Fukuyama et. al* in [20] and [4]. Compactness and separation are two internal

criteria proposed for evaluation of clustering. These criteria can also be used for selection of an optimal clustering scheme [2]:

1. Compactness: the members of each cluster should be as close to each other as possible. A common measure of compactness is the variance, which should be minimized.
2. Separation: the clusters themselves should be widely spaced.

There are three common approaches to measure the distance between two different clusters:

(i) Single linkage: Measures the distance between the closest members of the clusters.

(ii) Complete linkage: Measures the distance between the most distant members.

(iii) Comparison of centroids: Measures the distance between the centres of the clusters.

As mentioned before, two traditional internal cluster validity measures presented in [20], [4] employed in our experimental results (IV). These measure are as follows:

$$FS_m = \sum_{i=1}^N \sum_{j=1}^{nc} u_{ij}^m \left(\|x_i - v_j\|^2 - \|v_j - v\|^2 \right). \qquad (3)$$

Where,

$N :$ *number of data to be clustered*;
$nc :$ *number of clusters* ($nc \geq 2$);
$x_i :$ *i'th data*;
$v_j :$ *center of j'th cluster*;
$v :$ *average of all cluster centers*;
$\|\| :$ *norm*;
$u_{ij} :$ *grade of i'th data belonging to j'th cluster*;
$m :$ *adjustable weight* (*usually between* 1.5 *and* 3)

$$SC_m = \sum_{i=1}^N \sum_{j=1}^{nc} u_{ij}^m \left(\|x_i - v_j\|^2 - \|v_j - \overline{x}\|^2 \right) \qquad (4)$$

Where,

$N :$ *number of data to be clustered*;
$nc :$ *number of clusters* ($nc \geq 2$);
$x_i :$ *i'th data*;
$v_j :$ *center of j'th cluster*;
$\overline{x} :$ *average of data* $x_1, \; x_2, \; ..., x_n$;
$\|\| :$ *norm*;
$u_{ij} :$ *grade of i'th data belonging to j'th cluster*;
$m :$ *adjustable weight* (*usually between* 1.5 *and* 3)

Equations 3 [20] and 4 [4] consider both compactness and separation cluster validity criteria. The compactness part of the equations that is $\|x_i - v_j\|^2$ is the same. The SC separation term is based on the mean distance of datasets with the clusters that can be considered linkage based ($\|v_j - \overline{x}\|^2$). The FS separation term is based on comparison of centroids ($\|v_j - v\|^2$).

B. Trapezoidal Approximation

Trapezoidal approximation was employed to create membership functions (MF) for the ISSs (clusters) automatically.

1090

There are two algorithms that used grades of belonging in FCM for creating trapezoidal membership functions. The first work was presented by *Yasukawa and Sugeno* in [20]. It creates a convex hull over each of the clusters' degree of belonging values, and considers the best trapezoid which fits this convex hull as the MF for a sub-space.

The second algorithm was presented by *Tikk, Biro, Gedeon and Kóczy* [21]. In this algorithm the first step is to create a convex hull. After creating the convex hull, only the points that are located in corners of it will be used for the consequent steps of the algorithm. The next step is to select the point with highest degree of belonging to the current cluster (sub-space). By using this point, the algorithm divides the points into two groups and then creates the left and right slopes of the trapezoidal MF. More details of Tikk's trapezoidal approximation method can be found in [21] pages 598-599. After implementation of the second trapezoidal approximation, we came across an issue in the algorithm. To illustrate the issue we created a sample case in figure 1. In this figure, data points of a sample sub-space and their degree of belonging is presented. There are three data points with the highest value of 0.9 in the ISS. In figure 3 part (a) the convex hull that fit the data is drawn. Because the convex hull is polygonal and is defined by its corners, the middle point which is the optimal point for dividing the ISS to left and right slope is missing. Depend on implementation settings, *Tikk et. al.*'s trapezoidal algorithm will picked up either the first highest data point (figure 3 part (b))or the last one (figure 3 part (d)). In either of these cases the resulting membership function is not optimal for this ISSs. Selecting the first point as the maximum point to divide the points into left and right slopes, the resulting membership function in Tikk's trapezoidal algorithm is as it is shown in figure 3 part (c). For the other case the resulting membership function is illustrated in figure 3 part (e). For any other example which there exist several highest points and two of them are far from the center of the sub-space this problem will occur. In order to solve this problem we modified the first part of Tikk's trapezoidal algorithm. Instead of using the highest point on the convex hull to divide the ISS data into two slopes, we used center of gravity of data points. In our method:

1. The center of gravity of data points are calculated (figure 4 part (a))
2. Data points divide into two slopes
3. Tikk's algorithm is applied for each of the slope data (figure 4 parts (b) and (c)).
4. Finally the left and right slopes are merged to create the final trapezoidal MF (figure 4 parts (d)).

The issue of ISS data and their degree of belonging presented in figure 1 occures for sparse datasets. The reason is that for sparse datasets, the degree of belonging to each ISS is either high, close to one, or low, close to zero. As a result the possibility of having multiple equal highest values increases. In order to demonstrate the benefits of this modified version, we used 7 high salary selection problem [5] datasets with different levels of sparseness. *Gedeon et. al.* used dense fuzzy
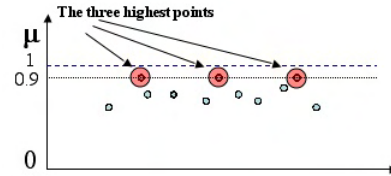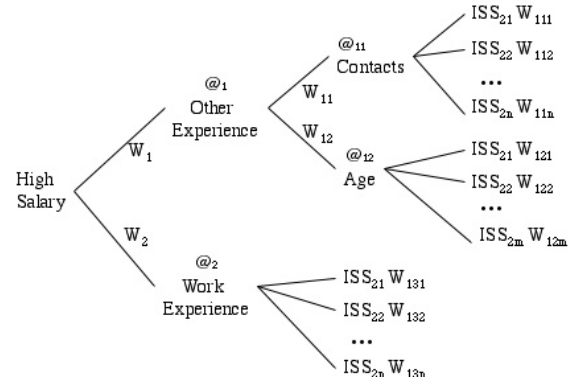

Fig. 1.   A sample sub-space of ISS

Fig. 2.   Structure of fuzzy signature for high salary selection problem

rule based system with 27 rules to model the salary problem. We used the same set of rules to generate the sparse datasets with different levels of spareness. For example, to create the first dataset in figures 6 and 7, we used only 15 randomly selected rules among 27 rules. The structure of the high salary selection PFS that is created by the automatic ISSs creation is presented in figure 2.

In this experiment, we considered only sparse datasets with middle range spareness (ie. from 15 -21) as datasets become too simple to approximate when only few rules were used to generate sparse data and if more rules were used to generate the data, the datasets were not as sparse. Experimental results in section IV, show that the modified trapezoidal approximation gave lower MSE for the middle range of sparse datasets, which are sparse as well as complex.

IV. MAIN RESULTS

There are two steps for creating ISS (clusters) for input data.

(i) Determining the optimal number of ISS for each input: FS (equation 3) and SC (equation 4) have been used.

(ii) Creating MFs based on degree of belonging for each ISS: Tikk's trapezoidal approximation and Center of Gravity (CoG) based trapezoidal approximation have been used.

As a result of employing these methods, we find four different ISS for salary fuzzy signatures. There are two Mean Square Errors (MSE) for test and train datasets for each of these four PFSs. We compared Tikk's trapezoidal approximation and CoG based trapezoidal approximation MSEs within these groups for both test and training datasets.

TABLE I

FS CLUSTER VALIDITY MSE FOR TRAINING DATA: CASES WHERE
MODIFIED METHOD HAS HIGHER MSE VALUE THAN THE ORIGINAL
METHOD IN TABLES ARE PRESENTED BY ITALIC STYLE

| Rules No.	CoG	Tikk's
15	0.1630	0.1703
16	0.1585	0.1743
17	0.0847	0.0884
18	0.0527	0.0542
19	0.1338	0.1342
20	0.0939	0.1110
21	*0.0494*	*0.0387*

TABLE II

FS CLUSTER VALIDITY MSE FOR TESTING DATA

Rules No.	CoG	Tikk's
15	0.1633	0.1706
16	0.1588	0.1753
17	0.0848	0.0886
18	0.0527	0.0542
19	0.1341	0.1345
20	0.0942	0.1112
21	*0.0494*	*0.0388*

MSE for testing and training sparse datasets implies that for all the datasets in training and testing results, the MSE of our method is lower except for few datasets with lower sparseness (figures 6 and 7). For FS cluster validity measure, it occurred for the dataset that was generated by 21 rules for both test and train datasets (tables I and II). For SC cluster validity measure, it occurred for two training datasets that was generated by 19 and 20 rules (table III) and for one test dataset generated by 19 rules (table IV).

The reason is, Tikk's resulting MFs are broader than our method (figures 3 parts (c) and (e)) for cases like figure 1.

In all these datasets there are data missing due to sparseness of the dataset. Hence, broader MF can sometimes help creating a PFS with lower MSE where there is data missing between two sub-space that look like figure 1. A sample of this case is demonstrated in figure 5. *Tikk et. al.*'s algorithm gives two possible set of MFs based on its implementation setting (figure 5 part (b) and (c)). Getting lower MSE due to missing data can not be considered as Tikk's algorithm advantage. Because as overall results imply it is not a common case; furthermore, the algorithm is not aware of the missing points.

In order to handle the missing data in training set, one of the possible category of techniques is fuzzy interpolation techniques [1], [25], [24].

Number of the times our CoG trapezoidal approximation method gives lower MSE than Tikk's trapezoidal approximation method for FS and SC cluster validity measures is presented in table V. Overlay, in this set of experiments our method gave lower MSE 82% of the time (table V values average).
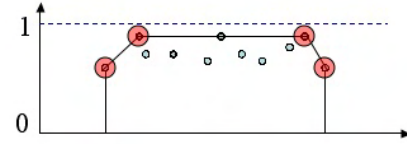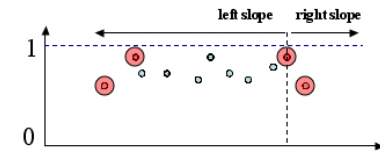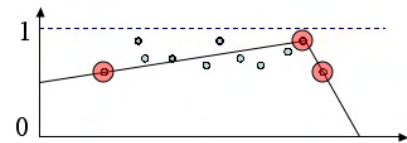

(a) data points convex hull
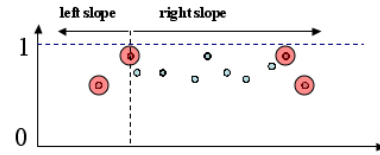
(b) Tikk's left and right slopes: First highest point is considered as the highest point
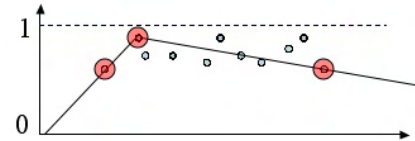
(c) Tikk's possible trapezoidal MF 1: First highest point is considered as the highest point

(d) Tikk's left and right slopes: Second highest point is considered as the highest point

(e) Tikk's possible trapezoidal MF 2: Second highest point is considered as the highest point

Fig. 3.   Tikk's trapezoidal approximation result for sample ISS

TABLE III

SC CLUSTER VALIDITY MSE FOR TRAINING DATA

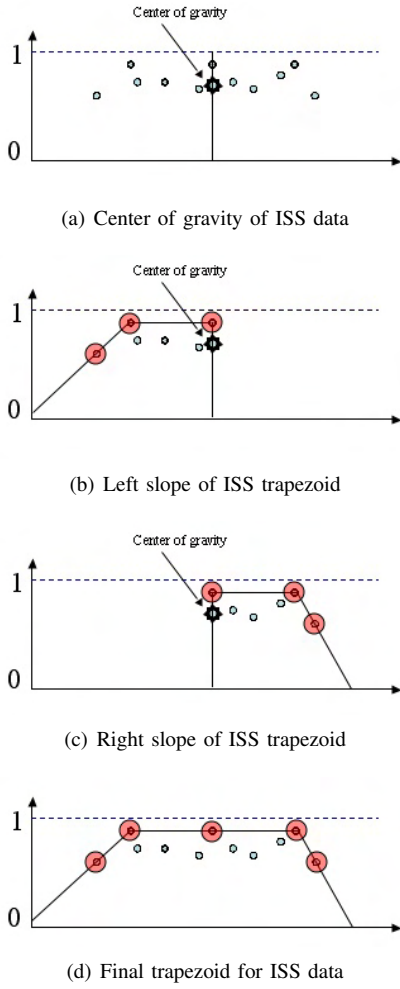| Rules No.	CoG	Tikk's
15	0.1520	0.1528
16	0.1632	0.1651
17	0.0729	0.0832
18	0.0384	0.0422
19	*0.1404*	*0.1387*
20	*0.1091*	*0.1089*
21	0.0368	0.0388

(a) Center of gravity of ISS data

(b) Left slope of ISS trapezoid

(c) Right slope of ISS trapezoid

(d) Final trapezoid for ISS data

Fig. 4.   Input dataset for a sample ISS

(a) Sample of ISS with missing points between two of its sub-spaces

(b) Tikk's algorithm resulting MFs in first highest point selection setting

(c) Tikk's algorithm resulting MFs in last highest point selection setting

(d) Our CoG based algorithm resulting MFs

Fig. 5.   Missing data points sample

TABLE IV

SC CLUSTER VALIDITY MSE FOR TESTING DATA

Rules No.	CoG	Tikk's
15	0.1523	0.1531
16	0.1635	0.1655
17	0.0730	0.0834
18	0.0385	0.0424
19	*0.1408*	*0.1390*
20	0.1093	0.1094
21	0.0369	0.0389

TABLE V

CoG AND TIKK COMPARISON FOR SC AND FS

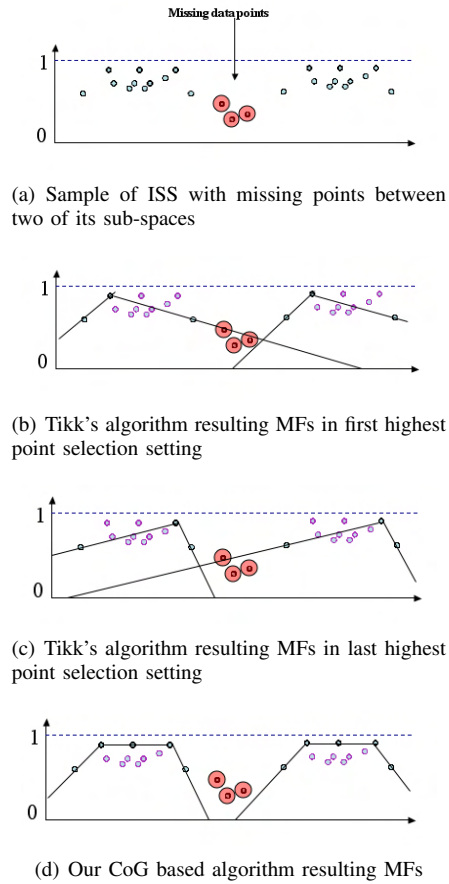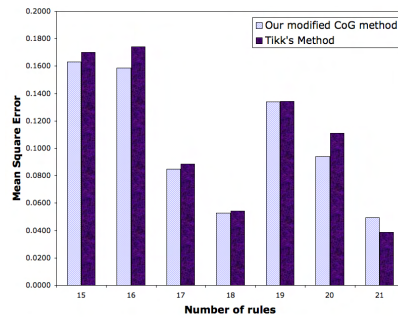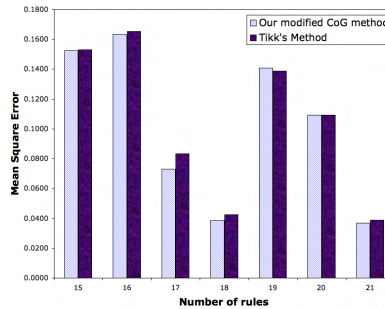|	FS	SC
Test	86%	86%
Train	71%	86%

(a) Train data MSE

(b) Test data MSE

Fig. 6.   FS measure data MSE; X-axis data is labeled in accordance with number of rules that are used to create the dataset
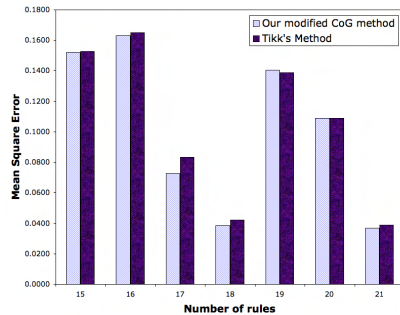
1093

(a) Train data MSE

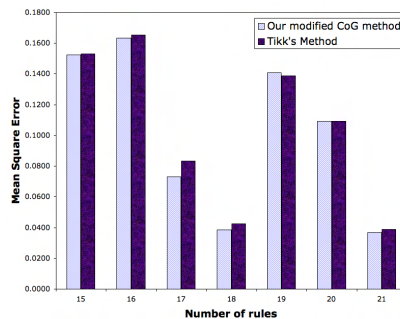
(b) Test data MSE

Fig. 7.   SC measure data MSE; X-axis data is labeled in accordance with number of rules that are used to create the dataset

V. CONCLUSION

A modified ISS finding has been presented which is based on CoG. By employing this method for fuzzy signatures, generating fuzzy signatures for sparse data is much simpler and more efficient than the manual approach.

In our experimental results we presented the MSE of PFS created by Tikk's and our CoG based trapezoidal approximation methods. This has been done for two cluster validity measures, SC and FS. For both cluster validity measures, overall results of training and testing data imply that our modified CoG based method has lower MSE than Tikk's method.

Future research, could include, employing parameter identification algorithm for ISS optimization [20] . This new PFS development approach employs not only fuzzy signatures weights and aggregations adjustments, but also ISS adjustments. This could lead to generation of a PFS with lower MSE.

REFERENCES

[1] P. Baranyi, L. Koczy, T. Gedeon, A generalized concept for fuzzy rule interpolation, Fuzzy Systems, IEEE Transactions on 12 (6) (2004) 820–837.

[2] M. Berry, G. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support, John Wiley & Sons, Inc. New York, NY, USA, 1997.

[3] J. C. Bezdek, R. Ehrich, Fcm: the fuzzy c-means clustering, COMP. GEOSCI. 10 (2) (1984) 191–203.

[4] Y. Fukuyama, M. Sugeno, A new method of choosing the number of clusters for the fuzzy c-mean method, In 5'th Proceeding of Fuzzy System Symposium (1989) 247–250.

[5] T. Gedeon, K. Wong, D. Tikk, Constructing hierarchical fuzzy rule bases for classification, Fuzzy Systems, 2001. The 10th IEEE International Conference on 3.

[6] M. Halkidi, Y. Batistakis, M. Vazirgiannis, Cluster validity methods: part i, ACM SIGMOD Record 31 (2) (2002) 40–45.

[7] J. Handl, J. Knowles, D. Kell, Computational cluster validation in postgenomic data analysis, Bioinformatics 21 (15) (2005) 3201–3212.

[8] A. Jain, R. Dubes, Algorithms for clustering data, Prentice Hall Advanced Reference Series, 1988.

[9] A. JAIN, M. MURTY, P. FLYNN, Data clustering: A review, ACM Computing Surveys 31 (3).

[10] L. Kóczy, Vector valued fuzzy sets, BUSEFAL- BULL STUD EXCH FUZZIN APPL (1980) 41–57.

[11] L. Kóczy, T. Vámos, G. Biró, Fuzzy signatures, in: EUROFUSE-SIC '99, 1999, pp. 210–217.

[12] D. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, Journal of the Society for Industrial and Applied Mathematics 11 (2) (1963) 431–441.

[13] B. S. U. Mendis, Fuzzy signatures: Hierarchical fuzzy systems and alpplications, Ph.D. thesis, Collage of Engineering and Computer Science, The Australian National University, Australia (2008).

[14] B. S. U. Mendis, T. D. Gedeon, Aggregation selection for hierarchical fuzzy signatures: A comparison of hierarchical owa and wrao, in: International Conference of Information Processing and Management of Uncertainty in Knowledge Based Systems (IPMU), Spain, 2008 (accepted).

[15] B. S. U. Mendis, T. D. Gedeon, A comparison: Fuzzy signatures and choquet integral, IEEE World Congress on Computational Intelligence, WCCI, Hong Kong.

[16] B. S. U. Mendis, T. D. Gedeon, L. T. Kóczy, Flexibility and robustness of hierarchical fuzzy signature structures with perturbed input data, in: International Conference of Information Processing and Management of Uncertainty in Knowledge Based Systems (IPMU), IPMU proceedings, Paris, 2006, pp. 2552–2559.

[17] B. S. U. Mendis, T. D. Gedeon, L. T. Kóczy, Generalised weighted relevance aggregation operators for hierarchical fuzzy signatures, in: International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA'06), Sydney, Australia, 2006, p. 198.

[18] B. S. U. Mendis, T. D. Gedeon, L. T. Kóczy, Learning generalized weighted relevance aggregation operators using levenberg-marquardt method, in: Sixth International Conference on Hybrid Intelligent Systems (HIS'06) and 4th Conference on Neuro-Computing and Evolving Intelligence (NCEI 06'), New Zealand, 2006, p. 34.

[19] B. S. U. Mendis, T. D. Gedeon, L. T. Kóczy, On the issue of learning weights from observations for fuzzy signatures, World Automation Congress (WAC).

[20] M. Sugeno, T. Yasukawa, A fuzzy-logic-based approach to qualitative modeling, In 5'th Proceeding of Fuzzy System Symposium 1 (1993) 7–31.

[21] D. Tikk, G. Biro, T. Gedeon, L. Koczy, Improvements and critique on sugeno's and yasukawa's qualitative modeling, Fuzzy Systems (2002) 596–606.

[22] K. Wong, A. Chong, T. Gedeon, L. Kóczy, T. Vámos, Hierarchical Fuzzy Signature Structure for Complex Structured Data, in: International Symposium on Computational Intelligence and Intelligent Informatics (ISCIII), Nabeul, 2003, pp. 105–109.

[23] K. Wong, T. Gedeon, L. Koczy, Construction of fuzzy signature from data: an example of SARS pre-clinical diagnosis system, in: Fuzzy Systems, 2004. Proceedings. 2004 IEEE International Conference on, vol. 3, 2004.

[24] K. Wong, T. Gedeon, D. Tikk, An improved multidimensional α-cut based fuzzy interpolation technique, in: Proc. Int. Conf Artificial Intelligence in Science and Technology (AISAT'2000), Hobart, Australia, 2000, pp. 29–32.

[25] K. Wong, D. Tikk, T. Gedeon, L. Koczy, Fuzzy rule interpolation for multidimensional input spaces with applications: a case study, Fuzzy Systems, IEEE Transactions on 13 (6) (2005) 809–819.

[26] K. W. Wong, T. Gedeon, L. Kóczy, Efficient fuzzy cognitive modeling for unstructured information, in: IEEE World Congress on Computational Intelligence, WCCI, Vancouver, Canada, 2006, pp. 358 – 363.

[27] K. Wu, M. Yang, A cluster validity index for fuzzy clustering, Pattern Recognition Letters 26 (9) (2005) 1275–1291.