PATTERN CLASSIFICATION FOR INCOMPLETE DATA

Bogdan Gabrys

Applied Computational Intelligence Research Unit
Department of Computing and Information Systems
The University of Paisley, High Street, Paisley PA1 2BE
Scotland, United Kingdom
Tel: +44 (0) 141 848 3752, Fax: +44 (0) 141 848 3542
E-mail: gabr-ci0@paisley.ac.uk

1. ABSTRACT

In this paper the problem of pattern classification for inputs with missing values is considered. A general fuzzy min-max (GFMM) neural network utilising hyperbox fuzzy sets as a representation of data cluster prototypes is used. It is shown how a classification decisions can be carried out on a subspace of a high dimensional input data. No substitution scheme for missing values (approach quite commonly used) is utilised. The result is a classification procedure that reduces a number of viable class alternatives on the basis of available information rather than attempting to produce one winning class without supporting evidence. A number of simulation results for well known data sets are provided to illustrate the properties and performance of the proposed approach.

2. INTRODUCTION

Missing values have been largely ignored in the pattern recognition literature, although problems with incomplete feature vectors are quite common and of great practical and theoretical interest [1, 4-8].

The reasons for missing data can be multifold ranging from sensor failures in engineering applications to deliberate withholding of some information in medical questioners. Whatever the reason for missing data the fact remains that most of the pattern recognition algorithms are not able to cope with such deficient inputs. It is in a sharp contrast to the very efficient way in which humans deal with unknown data and are able to carry out various pattern recognition tasks given only a subset of input features.

One of the most common ways of dealing with missing values is to substitute the missing features with their estimates [1]. These could include the mean value calculated over all examples (or k-nearest neighbours) for a particular feature. However, as it has been pointed out in [4,5,8] such a "repaired" data set may no longer be a good representation of the problem at hand and quite often leads to the solutions that are far from optimal. Instead the authors in the above mentioned papers used the approach based on estimating conditional probability distribution over all unknown features given the known features. And although it is another form of estimating the missing values it proved to be more accurate than the heuristic approaches used before.

Another approach presented in [7] advocates the creation of a set of classifiers that would work on different subsets of input features. Unfortunately, this method of training classifiers for all possible combinations of input features very quickly explodes in complexity with an increasing number of features.

One of the potential drawbacks of the methods based on the estimation of missing features is the fact that once the estimated value has been used (in implicit or explicit way) the output of the classifier does not differ in any way from the output generated for examples with all features present. To give an extreme example a classification of an input with no missing values can produce the same result as the classification of the input with all features missing. Clearly in the second case, since there is no evidence supporting the choice of any of the classes, this should be reflected in the classification output by making all the classes equally viable alternatives. In other words there are a number of applications (i.e. diagnostic analysis) where for inputs with missing features the aim of classification should be the reduction of viable alternatives rather than finding the most probable class. It is worth mentioning that the reduction of viable alternative classes does not rule out the possibility of choosing only one class as a result of classification as long as the evidence for this result is contained in the given subset of features.

In this paper the operation of the GFMM [2,3] neural network for classification of inputs with missing values is presented. The recall stage is only considered and the classification results are analysed. It is shown how a classification decisions can be carried out on a subspace of a high dimensional input data. A specific form of cluster prototypes (hyperbox fuzzy sets) in combination with a classification function based on a fuzzy membership function are shown to facilitate the classification of incomplete input vectors without any modifications done to the NN structure or substituting of missing features with estimated values. The result is a classification procedure that reduces a number of viable class alternatives on the basis of available information rather than attempting to produce one winning class without supporting evidence. The remaining of this paper is organized as follows. Section 3 presents a summary of the GFMM neural network with definitions of hyperbox fuzzy sets and associated fuzzy membership function. A way of dealing with

missing features within the GFMM NN structure is described in Section 4. The simulation results for a number of well known data sets follow in Section 5. And finally the conclusions are drawn in the last section.

3. AN OVERVIEW OF GFMM NEURAL NETWORK

The GFMM neural network for classification constitutes a pattern recognition approach that is based on hyperbox fuzzy sets. A hyperbox defines a region of the n-dimensional pattern space, and all patterns contained within the hyperbox have full class membership. A hyperbox is completely defined by its min-point and its max-point. The combination of the min-max points and the hyperbox membership function defines a fuzzy set. Learning in the GFMM neural network for classification consists of creating and adjusting hyperboxes in pattern space. For more details concerning the training algorithm please refer to [3]. Once the network is trained the input space is covered with hyperbox fuzzy sets. Individual hyperboxes representing the same class are aggregated to form a single fuzzy set class. Hyperboxes belonging to the same class are allowed to overlap while hyperboxes belonging to different classes are not allowed to overlap therefore avoiding the ambiguity of an input having full membership in more than one class. The input to the GFMM can be itself a hyperbox (thus representing features given in a form of upper and lower limits) and is defined as follows:

$$\boldsymbol{A}_h = [\boldsymbol{A}_h^l \;\; \boldsymbol{A}_h^u]$$

where $\boldsymbol{A}_h^l$ and $\boldsymbol{A}_h^u$ are the lower and the upper limit vectors for the *h*-th input pattern. Inputs are contained within the *n*-dimensional unit cube I^n. When $\boldsymbol{A}_h^l = \boldsymbol{A}_h^u$ the input represents a point in the pattern space.

The *j*-th hyperbox fuzzy set, B_j is defined as follows:

$$B_j = \{\boldsymbol{A}_h, \boldsymbol{V}_j, \boldsymbol{W}_j, b_j(\boldsymbol{X}_h, \boldsymbol{V}_j, \boldsymbol{W}_j)\} \qquad (1)$$

for all j=1,2,...,m, where $\boldsymbol{V}_j = (v_{j1}, v_{j2}, \ldots, v_{jn})$ is the min point for the *j*-th hyperbox, $\boldsymbol{W}_j = (w_{j1}, w_{j2}, \ldots, w_{jn})$ is the max point for the *j*-th hyperbox, and the membership function for the *j*-th hyperbox is:

$$b_j(\boldsymbol{A}_h) = \min_{i=1..n}(\min([1 - f(a_{hi}^u - w_{ji}, \gamma_i)], [1 - f(v_{ji} - a_{hi}^l, \gamma_i)])) \qquad (2)$$

where:

$$f(x, \gamma) = \begin{cases} 1 & if \quad x\gamma > 1 \\ x\gamma & if \quad 0 \le x\gamma \le 1 \\ 0 & if \quad x\gamma < 0 \end{cases} \text{ - two parameter ramp}$$

threshold function; $\gamma = [\gamma_1, \gamma_2, \ldots, \gamma_n]$ - sensitivity parameters governing how fast the membership values decrease; and $0 \le b_j(\boldsymbol{A}_h, \boldsymbol{V}_j, \boldsymbol{W}_j) \le 1$.

The membership values are used to decide whether the presented input pattern belongs to the class associated with the *j*-th hyperbox during the neural network operation stage.

The neural network that implements the GFMM classification algorithm is a three layer feedforward neural network shown at Fig. 1.

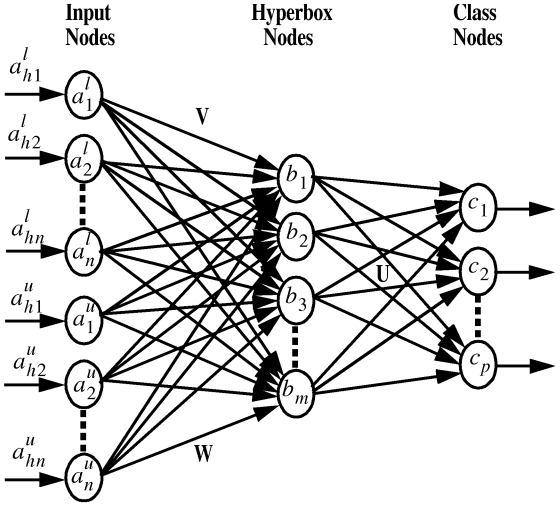

Figure 1: Three layer feedforward GFMM neural network.

The input layer has 2*n processing elements, two for each of the n dimensions of the input pattern $\boldsymbol{A}_h = [\boldsymbol{A}_h^l \;\; \boldsymbol{A}_h^u]$. Each node in the second layer represents a hyperbox fuzzy set. The connections of the first and second layer are the min-max points and the transfer function is the hyperbox membership function. The min points matrix **V** is applied to the first n input nodes representing the vector of lower bounds $\boldsymbol{A}_h^l$ of the input pattern, and the max points matrix **W** is applied to the other n input nodes representing the vector of upper bounds $\boldsymbol{A}_h^u$ of the input pattern. Each node in the third (output) layer represents a class. The connections between the nodes of the second and third layer are binary values assuming 1 if the second layer hyperbox fuzzy set is a part of the class represented by the output layer node and 0 otherwise. They are stored in the matrix **U**. The equation for assigning the values of **U** is

$$u_{jk} = \begin{cases} 1 & \text{if } b_j \text{ is a hyperbox for class } c_k \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

where b_j is the *j*-th node in the second layer and c_k is the *k*-th node in the third layer. The output of the third layer node represents the degree to which the

input pattern A_h fits within the class k. The transfer function for each of the third layer nodes is defined as

$$c_k = \max_{j=1}^{m} b_j u_{jk} \quad (4)$$

for each of the p third layer nodes. The outputs of the class layer nodes can be fuzzy when calculated using equation (4) directly, or crisp when a value of 1 is assigned to the node with the largest c_k and 0 to the other nodes.

4. INCOMPLETE DATA PROCESSING

It is assumed at this stage that the GFMM neural network has been trained to classify n-dimensional input patterns to one of the p classes. The classification is based on a distance measure given by the hyperbox membership function (2) and the hyperbox aggregation formula (4).

The discriminative character of the hyperbox membership function is based on penalising the violations of hyperbox min and max values for each input dimension. The smaller a_{hi}^l than v_{ji} or the larger a_{hi}^u than w_{ji}, the smaller the membership value, $b_j(A_h)$, for the j-th hyperbox fuzzy set B_j. If the i-th feature (dimension) of the input pattern is completely missing one would like to make sure that the hyperbox membership values will not be decreased due to this fact. It can be assured by the following assignments for the missing i-th feature:

$$a_{hi}^l = 1 \text{ and } a_{hi}^u = 0 \quad (5)$$

In other words all values for i-th feature are considered as equally likely as illustrated at Fig. 2 showing a simple 2 dimensional example.

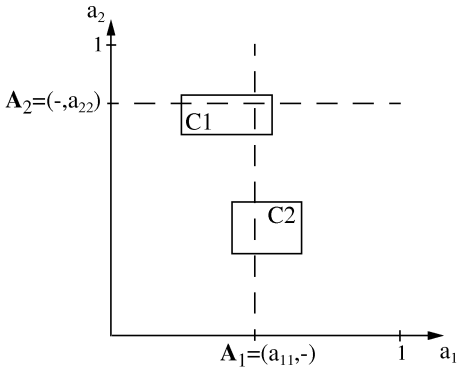

Figure 2: A simple 2 dimensional illustration of incomplete data processing using GFMM NN for two hyperbox fuzzy sets representing 2 classes (C1 and C2)

Case 1 - input pattern with second feature missing $\mathbf{A}_1=(a_{11},-)$; Result: $b_1(\mathbf{A}_1)=b_2(\mathbf{A}_1)=1$ - two equally likely alternatives.

Case 2 - input pattern with first feature missing $\mathbf{A}_2=(-,a_{22})$; Result: $b_1(\mathbf{A}_2)=1 > b_2(\mathbf{A}_2)$ - class C1 selected as a winner.

Assignment (5) also ensures that the structure of the NN will not have to be changed when processing inputs with missing dimensions. Another consequence of (5) is a possibility of an input with missing features having a full membership in more than one class. This, however, is something rather desirable since it aids in distinguishing from cases with all features present and reflects uncertainty associated with missing data. Generally, the more features missing the closer the membership value to 1 for more classes.

5. SIMULATION RESULTS

The testing of the proposed approach has been carried out on three well known data sets, namely IRIS, Wine and Ionosphere, obtained from the repository of machine learning databases (http://www.ics.uci.edu/~mlearn/MLRepository.html). The sizes and splits for testing and training data for all three data sets are shown in Table 1.

Data set	No. of inputs	No. of classes	No. of data points		
			Total	Train	Test
IRIS	4	3	150	75	75
Wine	13	3	178	90	88
Ionosphere	34	2	351	200	151

Table 1: The sizes of data sets used in classification experiments.

First the GFMM neural network was trained using training data sets without missing values. Subsequently, the testing was conducted on all testing data sets for a factor of missing features ranging from 10% to 80%. At each level, the average values over 100 testing runs, each for randomly chosen missing features have been calculated. The results obtained using the proposed method and their comparison with the nearest neighbour classifier, with missing values substituted with the mean values calculated from the training set, are shown at Fig. 3.

The greyed area represents cases where 2 or more classes have been identified as viable alternatives (classes with equal, maximal degree of membership) with the correct class always present. The area below the greyed area refers to unique, correct classification where only one, correct class has been chosen. The area above the greyed area represents misclassified cases.

What is interesting is the fact that the level of misclassified cases remains roughly the same for a whole range of missing features. The increasing level of missing features is reflected in a higher percentage of cases for which classification resulted in producing more than one viable alternative but almost always including the correct class. This can be particularly observed for IRIS and Wine data sets.

However, as it can be seen for Ionosphere data set

even for a high level of missing features, one winning class is produced for a vast majority of testing cases (i.e. there is a very small grey area). This is an example of how the classification with missing features can result in a one class being selected if only the known features are discriminative enough to do so.

Another interesting point is that the percentage of cases with multiple alternatives seem to increase more quickly with increasing ratio of missing features for data sets with smaller input data dimensionality. However, this observation will require further investigations and the connection between input data dimensionality and number of classes will have to be analysed more thouroughly.

Results for IRIS data presented in [4,5,8] fall within the grey area with a significant decrease in performance for higher percentage of missing values.

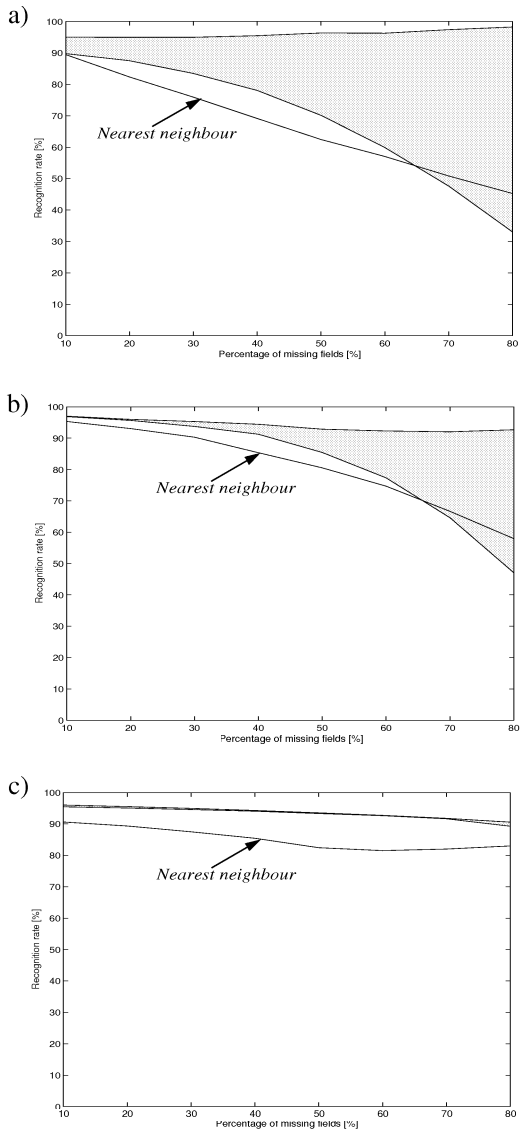

Figure 3: Classification results for a) IRIS, b) Wine, and c) Ionosphere data sets for different levels of missing features. Comparison of the proposed method with the nearest neighbour algorithm.

6. CONCLUSIONS

An approach based on GFMM neural network for dealing with incomplete data in classification problems has been presented. The resulting classifier uniquely classifies inputs with missing features if the evidence is present in the known features and produces a number of equally viable alternatives otherwise. In contrast to approaches based on estimating missing values or conditional probability distributions, uncertainty associated with missing features is directly reflected in the GFMM classifier's output. It stems from a more general property of the GFMM NN for classification and clustering which is able to distinguish between "equally likely" and "unknown" inputs [3].

When a classification process is required to produce a winning class the proposed method could be combined with approaches based on estimation of conditional probability distributions and thus providing more informative classification results.

REFRENCES

[1] J.K.Dixon, "Pattern Recognition with Partly Missing Data", *IEEE Transactions on Sys., Man, and Cyber.,* vol. SMC-9, no. 10, pp.617-621, October 1979

[2] B.Gabrys, A.Bargiela, "Neural Networks Based Decision Support in Presence of Uncertainties", *ASCE J. of Water Resources Planning and Management*, Vol. 125, No. 5, pp. 272-280, 1999

[3] B.Gabrys, A.Bargiela, "General Fuzzy Min-Max Neural Network for Clustering and Classification", accepted for publication in *IEEE Transactions on Neural Networks*, 2000

[4] Z.Ghahramani, M.I.Jordan, "Supervised Learning from Incomplete Data via an EM Approach", In: Cowan, J.D., Tesauro, G., and Alspector, J., eds., *Advances in Neural Information Processing Systems 6,* San Mateo, CA, Morgan Kaufman, 1994

[5] M.J.Nijman, H.J.Kappen, "Symmetry Breaking and Training from Incomplete Data with Radial Basis Boltzmann Machines", *International Journal of Neural Systems,* vol. 8, no. 3, pp. 301-315, 1997

[6] J.Kittler, "Classification of Pattern Vectors Using Modified Discriminant Functions", *IEEE Transactions on Computers,* vol. C-27, no. 4, pp. 367-375, April 1978

[7] P.K.Sharpe, R.J.Solly, "Dealing with Missing Values in Neural Network-Based Diagnostic Systems", *Neural Computing & Applications,* vol. 3, pp.73-77, 1995

[8] V.Tresp, R.Neuneier, S.Ahmad, "Efficient Methods for Dealing with Missing Data in Supervised Learning", In: G.Tesauro, D.S.Touretzky and K.Leen, eds., *Advances in Neural Information Processing Systems,* MIT Press, Cambridge MA, 1995