Unsupervised Feature Selection in high dimensional spaces and uncertainty

José R. Villar[1], María R. Suárez[1,1], Javier Sedano[2], Felipe Mateos[3],

[1] Computer Science Department, University of Oviedo, Campus de Viesques s/n 33204 Gijón, Spain
{villarjose, mrsuarez}@uniovi.es
[2] Electromechanic Engineering Department, University of Burgos, Spain
jsedano@ubu.es
[2] Electric, Electronic, Computers and Systems Engineering Department, University of Oviedo, Campus de Viesques s/n 33204 Gijón, Spain
felipe@isa.uniovi.es

Abstract. Developing models and methods to manage data vagueness is a current effervescent research field. Some work has been done with supervised problems but unsupervised problems and uncertainty have still not been studied. In this work, an extension of the Fuzzy Mutual Information Feature Selection algorithm for unsupervised problems is outlined. This proposal is a two stage procedure. Firstly, it makes use of the fuzzy mutual information measure and Battiti's feature selection algorithm and of a genetic algorithm to analyze the relationships between feature subspaces in a high dimensional space. The results of the first stage are used in the second with the aim to extract the most relevant relationships. It is concluded, given the results from the experiments carried out in this preliminary work, that it is possible to apply frequent pattern mining or similar methods in the second stage to reduce the dimensionality of the data set.

Keywords: Unsupervised feature selection, genetic algorithms, data uncertainty, frequent pattern mining.

1 Introduction

Many real world applications include a high dimensional feature space. Moreover, it is well known that the data gathered from a real world process could contain uncertainty [13], that is, there could be missing data, the measures could be interval values, etc. Typically the uncertainty in the data has been nullified by means of crisp techniques, i.e. the different techniques to eliminate missing data. What we are really doing is losing information about the process, and this information could be relevant in decision processes or in association rule discovering, especially in unsupervised

[1] Corresponding author: José R. Villar, Computer Science Department, University of Oviedo, Campus de Viesques s/n 33204 Gijón (Spain). E-mail: villarjose@uniovi.es.

problems, which represent an effervescent research topic due to its scarcity in the reported techniques [4].

On the other hand, high dimensional feature space represents a big challenge as a reduced data set is needed in order to reduce the over fitting of the models to be obtained. Also, high dimensional feature spaces increase the computational time needed in modeling such problems. Several different techniques have been employed to reduce the dimension of the data sets; they are known as feature reduction techniques and are divided into two main types: the feature extraction and feature selection techniques [10, 4]. Feature extraction includes the techniques that involve transforming the feature space into a smaller one. The transformation comprises any linear or nonlinear combination of a feature subset. An example of this kind of techniques is feature extraction by means of Principal Component Analysis [15].

Feature selection includes any method that proposes a feature subset from the original data set without any kind of transformation. The reduced feature space is supposed to include most relevant features according to a certain measure.

In this work, a feature selection technique able to deal with the data uncertainty is detailed. It is based on the Fuzzy Extension of the Mutual Information measure presented in [13], and it is designed for unsupervised problems. A two stages algorithm overcomes the problem of the dimensionality of the original data set. The new algorithm –called Fuzzy Unsupervised Mutual Information Feature Selection, from now on referred to as FUMIFS– has been found valid compared with previous approaches and some conclusions to improve the second stage have been extracted.

This work is organized as follows. A brief review of the feature selection methods and data uncertainty is outlined in the following section. In Section 3 the FUMIFS method is detailed. Section 4 deals with the experiments run and commented results. Finally, conclusions and future work is presented.

2. Uncertainty and feature selection in unsupervised problems

There are several feature selection techniques available in the literature. According to how the method must be used, feature selection methods are classified as *filters* or as *wrappers* [10, 17]. A feature selection method is referred to as a filter method if it is designed as a prepossess method before the modeling algorithm, i.e. [13]. When the feature selection is ran within the modeling algorithm then it is referred to as a wrapper method, i.e. the SSGA method [3]. The former methods are usually faster than the latter, with lower computation costs. In general, the performance of the wrapper methods is better than that of the filter methods, especially if the model obtained will be used to model the problem. Therefore, the wrappers are essentially designed for supervised problems.

According to how the method searches the domain, there are three possibilities: the *Complete Search* methods, the *Heuristic Search* methods and the *Random Search* methods. Also, the search is known as *Sequential Forward Search* -from now on, SFS- or *Sequential Backward Search* -from now on, SBS-. A heuristic search is called SFS if initially the feature subset is empty, and in each step it is incremented in one feature, i.e. the Battiti method [2]. On the other hand, it is an SBS if at the beginning

the feature subset is equal to the feature domain, and in each step the feature subset is reduced in one feature, i.e. the Fisher algorithm [12].

Although there are quite a lot of feature selection contributions reported in the literature, they are mainly designed for supervised problems [3, 17]. Moreover, the uncertainty included in the data is avoided in all of them, only crisp data is considered. Some unsupervised feature selection methods are also reported in the literature. In [5] the threshold that maximizes the mutual information is used in an SFS, choosing the features with higher mutual information values. Mitra et al proposed clustering feature subsets with the so-called maximum information compression index and choosing the most compact feature from each cluster [11]. Despite the speed and performance of the algorithm, the method is only designed for crisp data. Li et al proposed a hybrid method including a filter stage –using the fuzzy feature evaluation index– and a wrapper stage –using feature clustering. Finally, an unsupervised feature ranking is detailed in [7], where a ranking of the features is calculated based on clustering feature subsets, which they refer to as multiple view. For generating the feature subsets they proposed the random subspace method [6].

Imprecision and vagueness in data have been included in feature selection for modeling problems. Fuzzy logic has been employed for such task in the Fuzzy extension of the Mutual Information measure, which has been used in [13] to extend Battiti's algorithm for data uncertainty. Perhaps the rough set theory is the most widely used technique [9], all of them for supervised problems. A review of the rough set theory and the dimensionality reduction can be obtained in [16]. An SFS feature selection method for unsupervised problems using the neighborhood rough set is detailed in [8], where a neighborhood matrix is used to choose the features that maximize the neighborhood dependency in an SFS like algorithm. This feature selection method is specially defined to accomplish with heterogeneous data sets, that is, data sets that include both real valued and discrete valued features.

Some drawbacks should be commented. The majority of the feature selection methods in the literature do not consider uncertainty in the data and are mainly prepared for supervised problems. To our knowledge, only the last mentioned work deals with unsupervised problems using rough set theory. In general, it has been found that the performance of the SFS methods gets worse with the dimension of the domain space. Particularly, as the FMIFS is an SFS method for supervised problems that uses fuzzy theory, it is also concerned with this drawback. Finally, an increase in the computational cost has to be considered to manage uncertainty in the data.

3 The Fuzzy Unsupervised Mutual Information Feature Selection algorithm

In previous work an extension of Battiti's mutual information based feature selection method was proposed [13]. This extension, called Fuzzy Mutual Information Feature Selection –for short, FMIFS–, makes use of the fuzzy mutual information measure in order to deal with the uncertainty in the data. The robustness of the FMIFS performance against data uncertainty as missing data or interval-valued features

within the dataset was shown. Unfortunately, the FMIFS also shows the above-mentioned drawbacks [14].

The FUMIFS is proposed to overcome these disadvantages. In the FMIFS, the fuzzy mutual information measure is used to establish the information relationship between each variable and the class feature. In each step the feature with the highest value of residual mutual information of the feature class was included in the feature subset. This last step is dependant of a real value parameter called β, which represents the way the residual information of the feature class is calculated according to Eq. 1, where S is the set of features already chosen –that is, the best valued features subset– f is a feature in the domain that has not been chosen and C is the class feature.

$$RMF(f,C) = MF(f,C) - \beta \oplus_{sf \in S} MF(f,sf) \quad (1)$$

It has been found that the value of β is critical and problem dependant [14]. Moreover, when the number of features increases the residual mutual information is more influenced by the noisy variables. In such cases, the feature subset would include random variables, which are not related with the features in S. Nevertheless, if the number of features is relatively low the FMIFS behaves properly and it is a relatively fast method. Finally, the FMIFS is designed for supervised problems as reflected in Eq. 1 with the class feature C. Hence, the FUMIFS should exploit the behavior of the FMIFS when faced with relatively low dimension feature sets and must try to eliminate the influence of the random variables in high dimension feature sets. Also, as its main goal, the FUMIFS should manage unsupervised problems.

3.1 The unsupervised algorithm

The FUMIFS is based on some different approaches found in the literature. Firstly, the random sub-space method [6], which was also employed in [7], is used to choose a feature subset of lower dimension where the FUMIFS is intended to behave properly. The random sub-space method is applicable provided that there is no possibility of repeating the feature subset evaluation, which is to say that it should avoid evaluating a feature subset if it has already been evaluated. Secondly, a genetic algorithm is responsible for generating new feature subsets and evaluating them considering the restriction of the random sub-space method. So the individual is ranked according to how different it is compared to all the previously examined random subspaces, which in fact is the genetic fitness function. If an individual is found repeated then it is eliminated and a new one will be proposed.

Let N be the size of the random subspace. For each individual the FMIFS is run N times; in each run the feature from the random subspace used as objective feature is changed. Therefore, the K most relevant features according to the FIMFS are found for each feature in the feature subset. The value K represents the dimension of the feature subset proposed by the FMIFS in each run.

The individuals in the population are sorted according to their fitness. The genetic selection is carried out choosing individuals from the population with a probability

that decreases with the position in the sorted population. The crossover follows a two points crossover schema: according to the crossover probability the vector of included features of both parents are swapped to generate the two offsprings provided no repeated feature is included. In this case, the offspring is completed with a random chosen feature. The mutation goes through the vector of features of the individual to be mutated, and randomly changes each feature with the mutation probability. The vectors of features are always sorted according to their position in the original dataset.

When a population is completed then the certainty table is updated. The *Certainty Table* –for short, *CT*– accumulates the certainty that a feature depends on another. Each run of the FMIFS for an individual has an objective feature –for short, *of*– and proposes a K dimensional vector –for short, *vf*– of the most relevant features according to their mutual information measure. Then the certainty table is updated by means of Eq. 2, Eq. 3 and Eq. 4. Each value in CT is an interval value initialized to the crisp value of 0.

$$a_i = \min(CT(vf[i],of).\min(),\frac{1}{i}) \quad \forall i = 1 \cdots K \tag{2}$$

$$b_i = \max(CT(vf[i],of).\max(),\frac{1}{i}) \quad \forall i = 1 \cdots K \tag{3}$$

$$CT(vf[i],of) = Interval(a_i,b_i) \quad \forall i = 1 \cdots K \tag{4}$$

Finally, some relationships are extracted from the CT given the following rules of thumb. Let {LOW, MEDIUM, HIGH} be the linguistic terms of a fuzzy variable, and Let be f_i and f_j a pair of features for which relationships are to be found. The linguistic rules used to extract the relationships are: "*if CT(f_i, f_j) is HIGH and CT(f_i, f_j) is HIGH then there exists an Equivalence between f_i and f_j*" and "*if CT(f_i, f_j) is HIGH and CT(f_i,f_j) is LOW then there exists a DEPENDENCE of f_j in f_i*". These rules are used to prove that the algorithm is valid; it could be easily improved using the frequent pattern matching or any other algorithm that outperforms these simple rules.

Both N and K are parameters given to the FUMIFS. If N is set to the dimension of the original data set then the FUMIFS behaves like the FMIFS. K is typically set to less than half the value of N. The value of β should also be given as a parameter so FMIFS could be executed. The number of iterations (nIter), the population size (popSize) and the crossover and mutation probabilities must also be given. Care must be taken in setting the FUMIFS parameters to avoid infinite loops in the genetic algorithm. As the random subspace method is used there should not be a repeated individual. To prevent such an occurrence the number of iterations and the population size are bounded to not search more than the possible combinations of feature subspaces.

4 Experiments and results

The FUMIFS is to be compared with the FMIFS in order to test its goodness. So the same experimentation carried out in [13] is to be repeated for the FUMIFS. The datasets are available in the KEEL Project [1]. To provide unsupervised datasets the class feature has been considered as an input feature. The datasets have been modified to introduce vagueness, and both versions, the crisp and the imprecise ones, have been tested. After a FUMIFS run two data sets are generated: the first with all the features for which a relationship has been found with the class feature and the second data set with only those features for which dependency of the class feature was found. Then FMIFS has been run to choose the same number of features as FUMIFS. The values in all output files from each run are crisp according to [13], using the central point to convert an interval into a crisp value. Due to the length of this work, neither the crisp data sets results nor the boxplot graphics have been included. However, the results are commented.

The same thirteen different fuzzy rule-learning algorithms have been considered, both heuristic and genetic algorithms-based. In all cases, the number of linguistic terms in each partition is set beforehand, and not optimized by the learning algorithm. The experiments have been repeated ten times for different permutations of the datasets (10cv experimental setup). The heuristic classifiers use weighted fuzzy rules: always 1 (H1), the same weight as the confidence (H2), differences between the confidences (H3, H4, H5), weights tuned by reward-punishment (RE) and analytical learning (AL). The genetic fuzzy classifiers are the Genetic selection of rules taken from HEU3 (GE), Michigan learning (MI) –with population size 25 and 1000 generations–, Pittsburgh learning (PI) –with population size 50, 25 rules each individual and 50 generations–, the Hybrid learning (HY) –same parameters as PI, macromutation with probability 0.8–, the Fuzzy Ababoost (AD) –less than 25 rules with a single consequent, fuzzy inference by sum of votes– and Fuzzy Logitboost (LO) –less than 10 rules with multiple consequents, fuzzy inference by sum of votes–.

In Table 1 the classification mean errors for the thirteen methods are shown with the imprecise data sets. For each data set and method four results are given: the FUMIFS with only the dependence relationships found for the class feature (fumifs_d), the FUMIFS with all the relationships found for the class feature (fmifs_r), and the FMIFS results with the same number of features (fmifs_d and fmifs_r, respectively). Comparing the results of the FMIFS and the FUMIFS, it can be seen that the classification mean error is quite similar in both cases: with the dependence relationships and with any relationship between input features and the class output. Also the experiments run with the crisp data sets produced analogous results. Although they could not be included due to space limitations, the statistics boxplot graphics showed that results are totally comparable and it can not be concluded which method is better.

5 Conclusions and future works

In this work an unsupervised feature selection method has been described. It makes

	germandn0				ionosphere				pima			
	fmifs_r	fumifs_r	fmifs_d	fumifs_d	fmifs_r	fumifs_r	fmifs_d	fumifs_d	fmifs_r	fumifs_r	fmifs_d	fumifs_d
H1	0.259	0.300	0.317	0.300	0.126	0.140	0.204	0.129	0.314	0.140	0.357	0.129
H2	0.286	0.300	0.288	0.300	0.177	0.209	0.240	0.194	0.338	0.209	0.361	0.194
H3	0.285	0.300	0.302	0.300	0.183	0.203	0.241	0.197	0.345	0.203	0.351	0.197
H4	0.285	0.300	0.302	0.300	0.183	0.203	0.241	0.197	0.345	0.203	0.351	0.197
H5	0.285	0.300	0.302	0.300	0.103	0.203	0.206	0.197	0.345	0.203	0.351	0.197
RE	0.272	0.300	0.282	0.300	0.129	0.143	0.205	0.154	0.305	0.143	0.312	0.154
AL	0.274	0.300	0.288	0.300	0.137	0.140	0.208	0.137	0.328	0.140	0.295	0.137
GE	0.279	0.300	0.293	0.300	0.251	0.126	0.299	0.157	0.325	0.126	0.355	0.157
MI	0.300	0.300	0.300	0.300	0.114	0.311	0.230	0.309	0.350	0.311	0.350	0.309
PI	0.286	0.300	0.300	0.300	0.089	0.183	0.219	0.203	0.333	0.183	0.350	0.203
H	0.288	0.300	0.300	0.300	0.143	0.143	0.219	0.203	0.330	0.143	0.350	0.203
AD	0.265	0.292	0.277	0.294	0.149	0.149	0.218	0.200	0.230	0.149	0.288	0.200
LO	0.271	0.273	0.273	0.285	0.143	0.134	0.211	0.194	0.229	0.134	0.266	0.194

Table 1. FMIFS and FUMIFS classification mean error with the German Credit, the Ionosphere and the Pima Indian cancer data sets when vagueness is introduced in the data.

use of the Fuzzy Mutual Information measure and Battiti's algorithm which, combined with a genetic algorithm, generates a new data set that is to be post processed. It is proposed to use a frequent pattern matching method, but for this work only two rules of thumb were used. Results show that the FUMIFS behaves similarly to the previous work FMIFS. The FUMIFS is really influenced by the simple rules of thumb used. Also, both the aggregation method and the generation of the so-called certainty table have not been optimized. Nevertheless, this unsupervised feature selection method behaves properly, and the results encourage the authors to apply frequent pattern matching in order to improve the goodness of the relationships found.

Acknowledgments. This work was funded by the Spanish Min. of Science and Technology, grants TIN2005-08036-C05-05 and TIN2007-67418-C03-03.

References

[1] J. Alcala-Fdez, L. Sanchez, S. Garcia, M.J. D. Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernandez and F. Herrera. KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems. *Soft Computing, http://10.1007/s00500-008-0323-y*, 2009.

[2] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, vol. 5(4), pages 537-550, 1994.

[3] J. Casillas, O. Cordon, M. J. D. Jesus and F. Herrera. Genetic feature selection in a fuzzy rule-based classification system learning process for high-dimensional problems. *Information Sciences*, vol. 136, pages 135-157, 2001.
[4] T. W. S. Chow, P. Wang and E. W. M. Ma. A New Feature Selection Scheme Using a Data Distribution Factor for Unsupervised Nominal Data. *IEEE Transactions on Systems, Man and Cybernetics - PART B: Cybernetics*, vol. 38(2), pages 499-509, 2008.
[5] C. O. Conaire and N. E. Connor. Unsupervised feature selection for detection using mutual information thresholding. In *Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, 2008.
[6] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20(8), pages 832-844, 1998.
[7] Y. Hong, S. Kwong, Y. Chang and Q. Ren. Consensus unsupervised feature ranking from multiple views. *Pattern Recognition Letters*, vol. 29(5), pages 595 - 602, 2008.
[8] Q. Hu, D. Yu, Z. Xie and J. Liu. Fuzzy Probabilistic Approximation Spaces and Their Information Measures. *IEEE Transactions on Fuzzy Systems*, vol. 14(2), pages 191-201, 2006.
[9] R. Jensen and Q. Shen. Fuzzy-rough sets assisted attribute selection. *IEEE Transactions on Fuzzy Systems*, vol. 1(15), pages 73-89, 2007.
[10] F. Marcelloni. Feature selection based on a modified fuzzy c-means algorithm with supervision. *Information Sciences*, vol. 151, 2003.
[11] P. Mitra, C. A. Murthy and S. K. Pal. Unsupervised Feature Selection using Feature Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24(3), pages 301-312, 2002.
[12] J. A. Roubus, M. Setnes and J. Abonyi. Learning fuzzy classification rules from labelled data. *Information Sciences*, vol. 150, pages 77-93, 2003.
[13] L. Sanchez, M. R. Suarez, J. R. Villar and I. Couso. Mutual Information-based Feature Selection and Fuzzy Discretization of Vague Data. *International Journal of Aproximate Reasoning, DOI: http://dx.doi.org/10.1016/*, 2008.
[14] L. Sanchez, J. R. Villar and I. Couso. *Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. EUSFLAT, Genetic Feature Selection for Fuzzy Discretized Data, 2008.
[15] J. Sedano, J. R. Villar, E. S. Corchado, L. Curiel and P. M. Bravo. The application of a two-step AI model to an Automated Pneumatic Drilling Process. *Accepted to be published in the International Journal of Computer Mathematics*, 2008.
[16] K. Thangavel and A. Pethalakshmi. Dimensionality reduction based on rough set theory: A review. *Applied Soft Computing*, vol. 9(1), pages 1 - 12, 2009.
[17] O. Uncu and I. Turksen. A novel feature selection approach: Combining feature wrappers and filters. *Information Sciences*, vol. 177, pages 449-466., 2007.