Using a fuzzy mutual information measure in Feature Selection for Evolutionary Learning

Javier Grande
Computer Science Department
University of Oviedo
Campus Viesques
33204 Gijón
j.grandegundin@gmail.com

María del Rosario Suárez
Computer Science Department
University of Oviedo
Campus Viesques
33204 Gijón
mrsuarez@uniovi.es

José R. Villar
Computer Science Department
University of Oviedo
Campus Viesques
33204 Gijón
villarjose@uniovi.es

Abstract

When high dimensional datasets are fed to obtain a classifier it is known that learning methods do not perform in a suitable way. The larger the amount of features the higher the complexity of the problem, and the larger the time consumed in generating the outcome -the classifier or the model-. Feature selection has been proved as a good technique for eliminating features that do not add information of the system. There are several different approaches for feature selection, but up to our knowledge there are not many different approaches when feature selection is involved with imprecise data and genetic fuzzy systems. In this paper, a feature selection method based on the fuzzy mutual information is proposed. The outlined method has proved to be valid for classifying problems when expertise partitioning is given, and it represents the base of future work with pure imprecise data.

1 Introduction

When attempting to generate a classifier or a model based from a dataset which is obtained from a real process, there are some facts that must be taken into account [18, 17]. On the one hand, the number of features in the dataset, and the number of examples as well, will surely be high. Furthermore, it is not known which of the features are relevant or not, nor the interdependency relations between them. On the other hand, the data obtained from real processes is vague data due to the precision of the sensors and transducers, the losses in A/D conversions, the sensitivity and sensibility of the sensors, etc.

It is well known that the former fact is alleviated by means of the feature selection techniques. There are several techniques in the literature facing such a problem. This feature selection must be carried out in such a way that the reduced dataset keeps as much information as possible about the original process. In other words, redundant features and features that do not possess information about the process are the ones to be eliminated [24]. However, in the feature selection process it must be taken into account that datasets from real processes are imprecise, so the feature selection decisions must be influenced by such vagueness [22].

It is important to point out that the data impreciseness affects the way in which the behaviour of each feature is managed. Fuzzy logic has been proved as a suitable technique for managing imprecise data [15, 16]. Whenever imprecise data is present fuzzy logic is going to be used in order to select the main features so the losses in information from real processes could be reduced [17].

This paper intends to evaluate different approaches for feature selection in datasets gathered from real processes. The approaches must be valid to be extended with the fuzzy

mutual information (from now on referred as FMI) measure detailed in [19, 22], so the final method would face imprecise data. In this paper it will be shown that using expertise partitioning, and a feature selection method based on the FMI measure, a suitable approach for solving classification problems will provided. In order to prove that idea, the experiments are to compare the error rate for several classifiers when feature selection is applied. Finally some ideas about future work using the FMI are proposed.

The paper is set out as follows. Firstly, a review of the literature is carried out. Then, a description of the developed algorithms is shown in Sec. 3. Experiments run and results are shown in Sec. 4. Finally, conclusions and future work are commented in Sec. 5.

2 Feature selection methods

Real processes generate high dimensionality datasets. In other words, the obtained datasets have an important number of input features, which are supposed to describe the desired output. In practical cases, some input features may be ignored without losing information about the output. This problem is called *feature selection*, and it intends to chose the smaller subset of input features that best describes the desired output [11]. Fuzzy systems are known to be suitable when it is necessary to manage uncertainty and vagueness. The uncertainty in the datasets will influence the feature selection methods, and the fuzzy classifiers and models to be obtained. Feature selection methods related to the problem of managing uncertainty in data will be analyzed below.

There are several feature selection techniques available in the literature. Some authors have proposed a taxonomy of the feature selection algorithms according to how the method must be used and how the method works [9, 25]. According to how the method must be used, feature selection methods are classified as *filters* or as *wrappers*. As filters they are known the feature selection methods that are used as a prepossess method. As wrappers they are known the feature selection methods that are embedded in the whole solution methods, that is, in classification, the feature selection method is included in the optimization method used. The former methods are usually faster than the latter, with lower computation costs. But the wrapper methods performance is usually better than filter methods, and a more suitable feature set is supposed to be selected.

The *Relief* and the *SSGA Integer knn method* are an example of each type of feature selection method. The *Relief* method is a filter method that uses the knn algorithm and the information gain to select the feature subset [8]. The *SSGA Integer knn method* [3], which is a wrapper method, makes use of a filter feature selection method and then a wrapper feature selection method for obtaining a fuzzy rule based classifier. This wrapper makes use of a genetic algorithm to generate a feature subset which is evaluated by means of a knn classifier. A similar work is presented in [29].

In any case, a wrapper can also be used as a filter, as shown in [13]. In this work, a predefined number of features is given. An optimization algorithm is used to search for the combination of features that give the best classification error rate. Two subsets of features with the same classification error rate are sorted by means of distance measure, which assesses the certainty with which an object is assigned to a class.

According to how the method works there are three possibilities: the *complete search* methods, the *heuristic search* methods and the *random search* methods. The complete search methods are employed when domain knowledge exists to prune the feature search space. Different approaches are known for complete search methods: the *branch & bound* approach, which is assumed to eliminate all the features with evaluation function values lower than a predefined bound, and the *best first search* approach, which searches the feature space until the first combination of features that produces no inconsistencies with the data is obtained.

Heuristic search methods are the feature se-

lection methods that searchs for a well suited feature set by means of a heuristic search method and an evaluation function. The heuristics used are simple techniques, such hill-climbing could be. 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.

In [1] a SFS Method is detailed. This method makes use of the mutual information between each feature and the class and the mutual information between each pair of features. In each step the best evaluated feature -the one with the highest former mutual information measure- is chosen to be a member of the feature subset if the value of the latter mutual information measure is lower than a predefined bound. A similar feature selection application is the one presented in [28].

Another SFS method is presented in [9], where the fcm clustering algorithm is used to choose the features. Based on the discrimination index of a feature with regard to a prototype of a cluster, the features with higher index value are included in the feature subset. Although it is not feature selection but rather feature weighting, in [26] a gradient based search is used to calculate the weight vector and then a weigthed FCM to obtain a cluster from data is used.

The search is *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. Finally, the random search methods are those that make use of a random search algorithm in determining the smaller feature subset. Genetic algorithms are typically employed as the random search method.

In [14] a SBS method is shown using the Fisher algorithm. The Fisher algorithm is used for discarding the lowest evaluated feature in each step. The evaluating function is the Fisher interclass separability. Once the feature subset is chosen, then a model is obtained by means of a genetic algorithm. Another SBS contribution is shown in [11]. An interval model for features could be admitted. In this paper, a FCM clustering is run, and each feature is indexed according to its importance. The importance is evaluated as the difference between the Euclidean distances of the examples to the cluster prototype with and without the feature. The larger the difference, the more important the feature is. Each feature is evaluated with a real value although features are considered interval.

In [25] a boosting of sequential feature selection algorithms is used to obtaining a final feature subset. The evaluation function for the two former is the root mean square error. The third method uses a correlation matrix as feature evaluation function. Finally, the latter uses as feature evaluation function the inconsistency measure.

Random search methods make use of genetic algorithms, simulated annealing, etc. The works detailed above [3, 29] could be considered of this type. Also the work presented in [23] makes use of a genetic algorithm to select the feature subset.

Imprecision and vagueness in data have been included in feature selection for modelling problems. In [20, 21, 6, 27] SBS feature selection methods have been presented taking into account the vagueness of data through the fuzzy-rough sets. In [20] foundations are presented, where in [21] the SBS algorithm is detailed. Finally, an ant colony algorithm is employed in [6, 7].

The same idea has been successfully reported for classification purposes in [27], using the particle swarm optimization algorithm. An important issue concerning the t-norms and t-co norms is analyzed in [2], where non convergence problems due to the use of the max t-co norm is reported. Also, a solution by means of the product t-norm and the sum t-co norm is proposed.

3 The implemented feature selection algorithm

This paper deals with feature selection for obtaining classifiers with imprecise and vague problems. Mutual information is the tool intended to be used because it helps to chose

the features that possess maximum information about the desired output. In order to use such a measure in feature selection for classification problems, the Battiti feature selection algorithm has been shown as a fast and efficient solution. But, to our knowledge, the Battiti approach has not been used in regression problems, so it should be extended. Also, when there is imprecision in the data, the mutual information defined for crisp data is not valid. In such problems, the mutual information measure employed should manage vagueness and imprecision.

Extending the Battiti algorithm to regression problems is not difficult if a discretization schema is taken into account and applied as a dataset preprocess stage. But managing imprecision is a more difficult problem. The mutual information measure must be defined to include the imprecision in calculations. To the best of our knowledge, in the literature there is not approach to feature selection that accomplishes with pure uncertainty data.

In [19, 22] a definition of the Fuzzy Mutual Information (from now on, FMI) measure is done, and an efficient algorithm for computing such measure is presented. It is applied to feature discretization, and results have shown two main ideas. Firstly, the fuzzy mutual information measure defined is a valid measure for both for discrete and imprecise data. Moreover, the result of the FMI measure is the same if discrete data is fed. And secondly, the discretization with such measure outperforms those obtained with similar methods using the classical mutual information definition.

Concluding, we propose that the feature selection algorithm proposed by Battiti [1] (in following, MIFS) could be extended to regression problems by means of a discretization preprocess stage. But also, we propose the use of the FMI measure instead of the classic mutual information measure used in the referred paper. The whole algorithm, then, is shown in Fig. 1 which will be referred to as FMIFS. It is worth notiicing that there are not too many differences with the original algorithm. Specifically, when crisp data is given for a classification problem, the algorithm performs as the Battiti algorithm.

```
Input:  D the imprecise dataset, F the
feature space,
        d the desired output
        k the number of feature to be
chosen
Output:  H the feature subset
Let f_j the j-th feature in the F,
j ∈ [1, |F|]
Set H= {}
For each f ∈F
   compute I(d, f), I is the fuzzy mutual
information measure
Find a feature f that is not dominated
over I(d,f)
   Let F=F−f
   Let H=H∪f
Repeat until |H| == k
   Compute I(f, h) if it is not
available, where f ∈ F, k ∈ H
   Select f ∈F that is nondominated
over I(d, f) − β Σ_{h∈H} I(f, h)
      Let F=F−f
      Let H=H∪f
The output is H
```

Figure 1: Feature Selection Algorithm in presence of imprecise data

4 Experiments and results

This section will analyze how the FMI based feature selection method behaves. Two more feature selection methods are used to test the validity of our proposal, both from those implemented in the KEEL project [12]. Specifically, the feature selection methods employed are the Relief and the SSGA Integer Knn methods. The dataset tested is the wine dataset about the chemical analysis of wines grown in a specific area of Italy, with 13 features, 3 class values and 178 examples.

Moreover, thirteen different fuzzy rule learning algorithms have been considered, both heuristic and genetic algorithm based. The heuristic classifiers are described in [5]: no weights (HEU1), same weight as the confidence (HEU2),differences between the confidences (HEU3, HEU4, HEU5), weights tuned by reward-punishment (REWP) and analyt-

	Relief	SSGA	MIFS	FMIFS
HEU1	0.500	**0.176**	0.323	**0.176**
HEU2	0.411	0.176	0.323	**0.117**
HEU3	0.235	**0.147**	0.264	**0.147**
HEU4	0.205	0.235	0.205	**0.176**
HEU5	0.176	**0.147**	0.176	0.176
REWP	0.088	**0.058**	0.117	0.088
ANAL	0.235	**0.088**	0.235	0.117
GENS	0.117	0.147	0.205	**0.088**
MICH	0.647	**0.176**	0.617	0.205
PITT	0.117	0.176	0.205	**0.088**
HYBR	0.176	0.117	0.176	**0.058**
ADAB	0.058	**0.000**	0.058	0.029
LOGI	0.058	**0.029**	0.058	0.088
best	0	8	0	7

Table 1: The average classification error after the 10 k fold cross validation of the different fuzzy rule-based classifiers after performing a feature selection, with the original MIFS algorithm and with the modified version proposed in this paper. The number of features selected is 5 features for all of the methods.

ical learning (ANAL). The genetic classifiers are: Selection of rules (GENS), Michigan learning (MICH) –with population size 25 and 1000 generations,– Pittsburgh learning (PITT) –with population size 50, 25 rules each individual and 50 generations,– and Hybrid learning (HYBR) –same parameters as PITT, macromutation with probability 0.8– [5]. Lastly, two iterative rule learning algorithms are studied: Fuzzy Ababoost (ADAB) –25 rules of type I, fuzzy inference by sum of votes– [4] and Fuzzy Logitboost (LOGI) –10 rules of type III, fuzzy inference by sum of votes– [10]. All the experiments have been repeated ten times for different permutations of the datasets (10cv experimental setup), and are shown in Table 1 and in Fig. 2. As can be seen, it can not be stated which of the methods SSGA or the FMIFS is better, and both are better than the Relief and MIFS, as expected.

5 Conclusions and future work

Experiments show that the FMIFS could be a valid feature selection method. When discrete data is present the selected features are suitable. But more experimentation is needed in order to find the kind of problem for which this method best fits. Also, imprecise datasets must be generated and tested, for which the fuzzy mutual information measure has been developed. Future works also include analysing which missing data must be processed, and how this measure could be used with different feature selection methods apart from that of Battiti. More work has to be done extending the proposed algorithm to obtain regression models.

Acknowledgment

This work was funded by Spanish Min. of Education, under the grant TIN2005-08386-C05.

References

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

[2] Bhatt, R. B., and Gopal, M. On fuzzy-rough sets approach to feature selection. *Pattern Recognition Letters*, 26 (2005), 965–975.

[3] Casillas, J., Cordón, O., Jesus, M. J. D., and Herrera, F. Genetic feature selection in a fuzzy rule-based classi®cation system learning process for high-dimensional problems. *Information Sciences*, 136 (2001), 135–157.

[4] del Jesús, M. J., Junco, F. H. L., and Sánchez, L. Induction of fuzzy-rule-based classifiers with evolutionary boosting algorithms. *IEEE Transactions on Fuzzy Systems 12*, 3 (2004), 296–308.

[5] Ishibuchi, H., Nakashima, T., and Nii, M. *Classification and Model-*

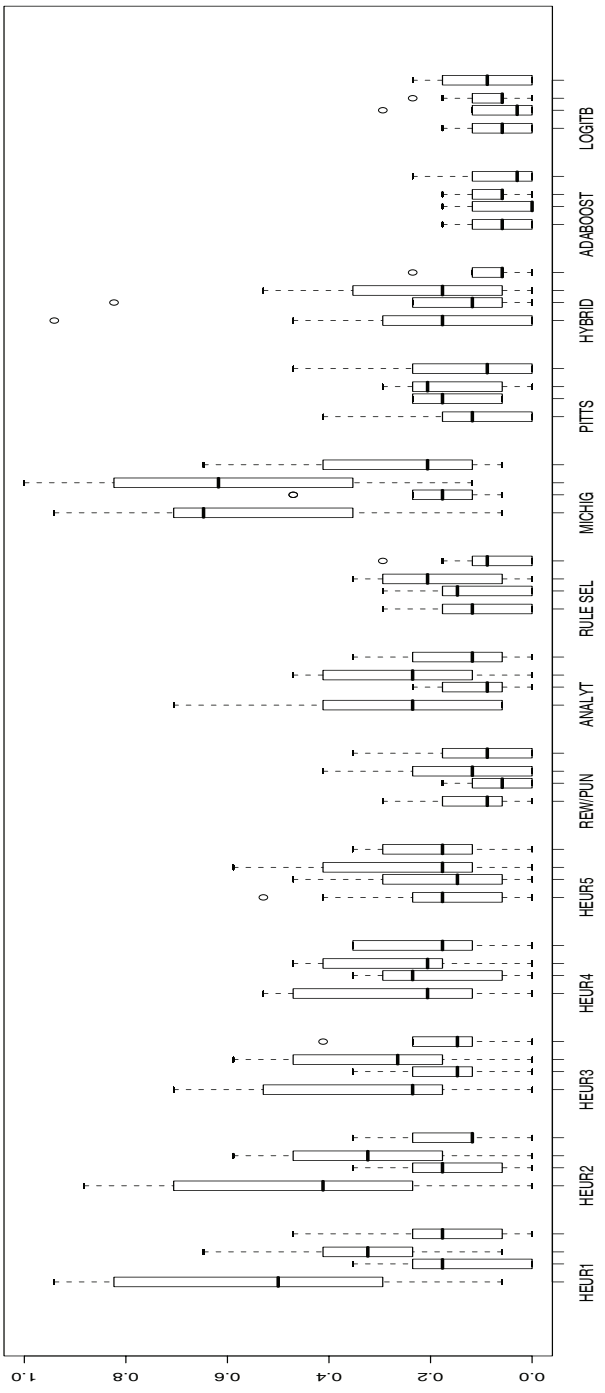

Figure 2: Boxplot for the error classification results after the 10 k fold cross validation for all of the classifier methods with the wine dataset.

ing with Linguistic Information Granules. Springer, 2004.

[6] Jensen, R., and Shen, Q. Fuzzy-rough data reduction with ant colony optimization. *Fuzzy Sets and Systems*, 149 (2005), 5–20.

[7] Jensen, R., and Shen, Q. Fuzzy-rough sets assisted attribute selection. *IEEE Transactions on Fuzzy Systems 15*, 1 (2007), 73–89.

[8] Kira, K., and Rendell, L. A practical approach to feature selection. In *Proceedings of the Ninth International Conference on Machine Learning (ICML-92)* (1992), pp. 249–256.

[9] Marcelloni, F. Feature selection based on a modified fuzzy c-means algorithm with supervision. *Information Sciences*, 151 (2003), 201–226.

[10] Otero, J., and Sánchez, L. Induction of descriptive fuzzy classifiers with the logitboost algorithm. *Soft Computing 10*, 9 (2005), 825–835.

[11] Pedrycz, W., and Vukovich, G. Feature analysis through information granulation and fuzzy sets. *Pattern Recognition*, 35 (2002), 825–834.

[12] Project, T. K. http://www.keel.es. Tech. rep.

[13] Ravi, V., and Zimmermann, H.-J. Fuzzy rule based classi®cation with featureselector and modified threshold accepting. *European Journal of Operational Research*, 123 (2000), 16–28.

[14] Roubos, J. A., Setnes, M., and Abonyi, J. Learning fuzzy classification rules from labeled data. *Information Sciences*, 150 (2003), 77–93.

[15] Sánchez, L., and Couso, I. Advocating the use of imprecisely observed data in genetic fuzzy systems. In *Proceedings of I International Workshop on Genetic Fuzzy Systems, GFS 2005* (2005).

[16] Sánchez, L., and Couso, I. Advocating the use of imprecisely observed data in genetic fuzzy systems. *IEEE Transactions on Fuzzy Systems, in press* (2006).

[17] Sánchez, L., Otero, J., and Casillas, J. Modeling vague data with genetic fuzzy systems under a combination of crisp and imprecise criteria. In *Proceedings of the First IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making, MCDM2007* (Honolulu, USA, 2007).

[18] Sánchez, L., Otero, J., and Villar, J. R. Boosting of fuzzy models for high-dimensional imprecise datasets. In *Proceedings of the Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU06* (Paris, France, 2006).

[19] Sánchez, L., Suárez, M. R., and Couso, I. A fuzzy definition of Mutual Information with application to the design of Genetic Fuzzy Classifiers. In *Proceedings of the IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2007* (London, UK, 2007).

[20] Shen, Q., and Chouchoulas, A. A rough-fuzzy approach for generating classiÿcation rules. *Pattern Recognition*, 35 (2002), 2425–2438.

[21] Shen, Q., and Jensen, R. Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring. *Pattern Recognition*, 37 (2004), 1351–1363.

[22] Suárez, M. R. *Estimación de la información mutua en problemas con datos imprecisos.* PhD thesis, University of Oviedo, Gijón, Spain, April 2007.

[23] Tolvi, J. Genetic algorithms for outlier detection and variable selection in linear regression models. *Soft Computing*, 8 (2004), 527–533.

[24] Tourassi, G. D., Frederik, E. D., Markey, M. K., and Carey E. Floyd, J. Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Med. Phys. 28*, 12 (2001), 2394–2402.

[25] Uncu, O., and Turksen, I. A novel feature selection approach: Combining feature wrappers and filters. *Information Sciences*, 177 (2007), 449–466.

[26] Wang, X., Wang, Y., and Wang, L. Improving fuzzy c-means clustering based on feature-weight learning. *Pattern Recognition Letters*, 25 (2004), 1123–1132.

[27] Wang, X., Yang, J., Jensen, R., and Liu, X. Rough set feature selection and rule induction for prediction of malignancy degree in brain glioma. *Computer methods and Programs in Biomedicine*, 83 (2006), 147–156.

[28] Yu, D., Hu, Q., and Wu, C. Uncertainty mearures for fuzzy relations and their applications. *Applied Soft Computing*, 7 (2007), 1135–1143.

[29] Yu, S., Backer, S. D., and Scheunders, P. Genetic feature selection combined with composite fuzzy nearest neighbor classiers for hyperspectral satellite imagery. *Pattern Recognition Letters*, 23 (2002), 183–190.