A Feature Selection method using a fuzzy mutual information measure

María del Rosario Suárez
Department of Computer Science,
University of Oviedo, Edif. Departamental 1, Campus de Viesques s/n, 33204
Gijón, Asturias, Spain
E-mail: mrsuarez@uniovi.es

José R. Villar *
Department of Computer Science,
University of Oviedo, Edif. Departamental 1, Campus de Viesques s/n, 33204
Gijón, Asturias, Spain
E-mail: villarjose@uniovi.es
*Corresponding author

Javier Grande
Software Engineering Department
Indra Sistemas, S.A., Moisés de León 57 24006, León, Spain
E-mail: j.grandegundin@gmail.com

Abstract: Attempting to obtain a classifier or a model from datasets could be a cumbersome task, specifically, when using datasets of high dimensionality. The larger the amount of features the higher the complexity of the problem, and the larger the time that is expended in generating the outcome -the classifier or the model-. Feature selection has been proved as a good technique for choosing features that best describes the system under certain criteria or measure. There are several different approaches for feature selection, but until 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 is valid for classifying problems when expertise partitioning is given, and it represents the base of future work including the use of the in case of imprecise data.

Keywords: mutual information; classification; feature selection; imprecise data; genetic fuzzy systems.

Reference to this paper should be made as follows: .

Biographical notes: M. R. Suárez received his PhD in Computer Science from University of Oviedo in 2007. She is currently Assistant Professor at the Department of Computer Science, University of Oviedo, Gijón, Spain. Her current research interest include managing process imprecise data, fuzzy systems and webmining.

J. R. Villar received his PhD in Computer Science from University of Leon in 2003. Ejerció profesionalmente en la ingeniería ISOTRON, S.A, en la Central Térmica de La Roble de la empresa eléctrica Unión FENOSA, S.A., en el instituto de investigación energética y minera ENERMITEC de León, y en la Universidad de León. He is currently Assistant Professor at the Department of Computer Science, University of Oviedo, Gijón, Spain. His current research interest include managing process imprecise data, fuzzy systems and intelligent industrial and real world applications under uncertainty.

J. Grande is with the Software Enginering Department at Indra Sistemas, S. A., León, Spain. He received his degree in Computer Science from University of León, Spain. His current research focuses on fuzzy systems and real world applications under uncertainty.

1 INTRODUCTION

When attempting to generate a classifier or a model based from a dataset 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 must be the chosen ones to be eliminated [24].

As stated before, the chosen subset must keep as much system information as possible. But information about the process is evaluated by means of a certain measure, which is known appropriated for some kind of problems. As a conclusion, a feature selection method, which makes use of a certain measure, will be suitable for some kind of problems, and could be outperformed by other methods under different circumstances. There is not a feature selection method that could accomplished correctly in all kind of problems.

However, the impreciseness in data must be taken into account in the feature selection process, 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 faces a rather common situation in genetic fuzzy systems, that is the use of a predefined fuzzy partition for all of the features. In this situation, the vagueness in the data introduces errors in the inference, and so in the output of a fuzzy model.

This paper intends to evaluate different approaches for feature selection in standard datasets. A feature selection method based in the mutual information measure is to be extended with the fuzzy mutual information measure (from now on referred as FMI) 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 be 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, and in Sec. 4 experiments ran and results are shown. Related work will be detailed then, and finally, conclusions and future work are commented.

2 AN OVERVIEW TO 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 choose the smaller subset of input features that best describes the desired output [11]. Unfortunately, the data from real processes are vague. Vagueness in data came in the form of loss data, the roundness of the samples in the analog to digital conversions, etc. The uncertainty in a dataset will influence the feature selection method outcome, but also in the 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 are 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 selection methods that search 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 fuzzy c-means (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 values 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 weighted 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. A method to manage vagueness in a feature selection method is by means of using a suitable measure that could take into account the uncertainty, as fuzzy systems could be [30]. In [20, 21, 6, 27, 28, 31, 32] 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.

In [31,32] a mixed feature selection method is proposed using the rough set that takes into account the different types of attributes –numerical or categorical –, and establishing lower and upper bounds of approximation, i.e., managing intervals of approximation of the decision. The proposed solution makes used only of the lower approximation of the decision in order to evaluate the feature subset.

As stated, there are several feature selection techniques that have been designed to be used with fuzzy systems, but up to our knowledge no feature selection method is designed to managed interval or fuzzy data.

3 THE FUZZY FEATURE SELECTION PROPOSAL

This paper deals with feature selection for obtaining classifiers with imprecise and vague problems. As seen in previous section, there are several different techniques for feature selection. This work first must choose a feature selection method and then extend it to imprecise data.

Mutual information is the tool intended to be used because it helps to choose 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 vagueness 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.

In following, our proposal is detailed. Firstly, a definition for imprecise data is given, and the relationship between the fuzzy partitions and the outcome of fuzzy models is analyzed. Then, the Battiti's filter feature selection method is detailed, and the extension of such algorithm to manage imprecise data as well.

3.1 A definition of imprecise data

In some problems, mainly in those involved with real processes, the information provided on the magnitudes involved is not accurate and includes uncertainty, i. e. for those magnitudes there aren't defined by a crisp value but by a imprecise value.

Impreciseness can be given in the form of noise in the data, or as the degree of precision of the measure, or the roundness of the measure. In case of data including noise statistical techniques deal with the random variables generated.

In the rest of cases, we do have inaccurate data where the probability distribution of the real value in not known, we are talking about a random set, which defines a family of random variables, and we could try to approximate the probability distribution of the real value.

We are interested in those numeric variables that are likely to become fuzzy sets from a fuzzy partition of the data. Depending on the fuzzy partition used, fuzzy set obtained will be different.

Given the linguistic partition in Figure 1, if the value 45 is given, the result we get is (0.0/COLD +0.2/WARM +0.8/HOT), where the sum of the probabilities is equal to 1. But if the value 45 ± 1 is given, the result we get is (0.0/COLD +0.3/WARM + 0.9/HOT), for which the sum of the probabilities is greater than one.

This is an imprecise data represented by means of an interval, and using a fuzzy partition of the feature universe. In this case, the state of no information could be given as (0.5/COLD + 0.5/WARM + 0.5/HOT).

3.2 Fuzzy partitioning issues and relevance

Partitioning a feature is a very important task, which is mainly carried out by an expert or by automated learning. When carried out through automated leaning the obtained partition represents the best one respecto to the objective function that was used. When the partitioning is obtained from an expert, then the partition could not be the best one with respect to a certain objective function, as it will be shown. Membership function is important in the feature selection and it has to be considered by the algorithm used.

In Figure 2 there we have the two variable two class problem. If we calculate the mutual information between the variable each and the class, we will have the variable X with a larger value than the variable Y. This is correct as the variable X gives enough class information to discriminate each possible value.

However, if we consider the partitioning scheme that is repsented in Figure 2, and we recalculate the mutual information between each variable and the class, we will found that the variable X variable behaves worse than the variable Y, and that although black spots are all low, the white dots can be low or high. As a conclusion, the measure use as objective function must not only deal with imprecise data but also with partitioning schemas. It is desirable that the impact of the partitioning schema has a relevance in the objective function behaviour as lower as possible.

As shown, the classical definition of mutual information is totally dependant of the partitioning schemas, so it must be extended. In [19, 22] we provide a new definition of Mutual Information in the case of inaccurate information.

In the case of inaccurate information we could have interval data. For an interval value, we should found the probability distribution of such data in the partitioning schema. Given this probability distribution we can calculate the mutual information. It is shown that a lower and an upper bounds for the probability distribution could be established. In the end, an interval mutual information value is obtained. If fuzzy data is given, then a fuzzy mutual information value is obtained, but if crisp data is given, then the result must be the same as with the classical mutual information measure. Interested readers should found all of the development in [19, 22].

3.3 Battiti feature selection algorithm

The mutual information feature selection (MIFS) method defined by Battiti in his seminal work [1] is one of the most widely used filter feature selection methods. The algorithm is outlined in Figure 3, which represents a SFS feature selection technique.

From the empty chosen features set, in each step the mutual information between each feature and the class is calculated, but also the mutual information between a feature in the dataset and each feature in the chosen subset.

Firstly, the feature with higher mutual information value is chosen. In each step, the feature with higher mutual information with the class and which is intended as the most independent from the rest of chosen features, that is, the information gain is used as objective function.

$$IG(Class, f) = MI(Class, f) - \beta \times \sum_{s \in S} MI(f, s)$$

A predefined constant factor β is used to weigh the relevance of the redundancies between chosen features. In the end, the algorithm will choose the K best valued features, and all of them are included in the chosen feature subset S.

3.4 Extending of the Battiti's algorithm to manage imprecise data

As shown in previous sections, the classical definition of mutual information is totally dependant of the partitioning schema. When using expert defined partitioning this is a not desirable behaviour, so this measure must be redefined for a better performance. In [19, 22] we provide a new definition of Mutual Information in the case of inaccurate information. This measure is known as Fuzzy Mutual Information measure (FMI).

In the case of inaccurate information we could have interval data. For an interval value, we should found the probability distribution of such data in the partitioning schema. Given this probability distribution we can calculate the mutual information. It is shown that a lower and an upper bounds for the probability distribution could be established. In the end, an interval mutual information value is obtained. If fuzzy data is given, then a fuzzy mutual information value is obtained, but if crisp data is given, then the result must be the same as with the classical mutual information measure. Interested readers should found all of the development in [19, 22].

If this new measure is to be used in the Battiti's feature selection algorithm, then it must be also modified, so it could managed not only crisp but imprecise data. In this sense, the grey zones shown in Figure 3 must be redesigned so they could accomplish with the new FMI. The new algorithm is the one reflected in Figure 4, in this case the grey zones are the correspondent modifications in the original algorithm.

Because of Battiti algorithm is a greedy algorithm, we began with an empty set of selected features and need select the first one. Let F be the features in the original dataset, and let |F| be cardinality of F. Let S be the chosen feature subset, initially empty. Let f_i be a feature in F, and s_i a feature in S.

Then, our first modification is when the algorithm calculates de mutual information between each feature and the class. The result are |F| mutual information values, N of them are interval values of mutual information, with N in [1, |F|]. Choosing the higher value induce the use of a relation of order, which must establish the rules to sort the values. For sorting the crisp and interval values it is needed to made interval comparisons based in dominance. Given two intervals A≡[a, b] and B≡[c, d], with a≤b and c≤d, then it is said that A dominates B ⇔ b≤c.

Once the first feature is chosen, the algorithm must select the next (K-1) features. In each step, the FMI must be calculated between all the possible pairs <f_i, s_i>. Then, the feature f_i with a higher information of the class is chosen. The information about the class that a feature has is calculated by means of information gain shown in the following equation, which is interval valued too, so interval arithmetic must be used.

$$IG(Class, f_i) = FMI(Class, f_i) \overset{\bullet}{-} \beta \times \sum\nolimits_{s_j \in S} FMI(f_i, s_j)$$

In each step, it is needed to sort the features in F with respect to the information gain in order to choose the best-valuated feature, which will be extracted from F and included in S.

It is important to mention that the extension of the MIFS to include the FMI (known as Fuzzy MIFS, FMIFS) should generate the same results than MIFS when crisp datasets are given.

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 are the german dataset is about a public benchmark from the UCI Machine Learning Repository, with 20 features, 2 class values and 1000 examples; the wine dataset is about the chemical analysis of wines grown in a specific area of Italy, with 13 features, 3 classes and 178 examples; the ion dataset is about data in the ionosphere with 34 features, 2 classes and 351 exmples; the pima dataset is about the Pima Indians Diabetes with 8 features, 2 class values and 768 examples; and the sonar dataset about the classification of sonar signals using a neural network, with 60 features, 2 class values and 208 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 analytical 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

Because of space reasons, we limit ourselves to crisp data and study the effect of including information about the fuzzy partition in the feature selection algorithm. In Figure

5, Figure 6, Figure 7, Figure 8 and Figure 9 we have compared the results of the new algorithm FMIFS for five crisp datasets to those of the original MIFS algorithm, the RELIEF [8] and the evolutionary algorithm SSGA [3]. In all cases, a uniform partition of size 3 was used for all the variables, and 5 input variables were selected. The algorithm SSGA was not different from the best one in 19 cases, while FMIFS was the best choice in 47 cases, SSGA in 30 cases, RELIEF in 8 cases and the crisp version of MIFS was the best in 6 cases. Observe that there are two problems were both FMIFS and SSGA improve the results of the crisp feature selection. In the remaining problems, the use of a fuzzy method did not degrade the results, and SSGA is not different than its crisp version.

Therefore, we think that this algorithm is a good compromise. In future works we will include compared results of the performance of SSGA and FMIFS over coarsely measured data and data with missing values.

5 RELATED WORK

Feature selection with imprecise data has not been studied in deep. There are some new and promising proposals, using different techniques. In one hand there are solutions based in rough sets with numerical data, on the other hand there are solutions based not only in numerical data but also in interval and fuzzy data. Feature selection techniques based on rough sets had been commented previously in the second section. Feature selection techniques dealing with imprecise data using interval and fuzzy data are those presented in [19, 33, 34, 35].

In [19, 33] a seminal work are presented. The fuzzy mutual information measure, and the advantages of its used are shown. In [34] the preliminaries of this work are given. Finally, in [35] a more on deep analysis and test bed for the mutual information based partitioning method and feature selection method are shown.

6 CONCLUSION 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 better fits. Also, imprecise datasets must be generated and tested, for which the fuzzy mutual information measure has been developed. Future works also includes analysing who missing data must be processed, and how this measure could be used with different feature selection methods apart from that of Battiti.

ACKNOWLEDGEMENT

This work was funded by Spanish M. of Education, under the grant TIN2005-08386-C05. Authors are thankful with Dr. Luciano Sánchez for his disinterested comments and the reviews of our work.

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., JESÚS, M. J. D., AND HERRERA, F. Genetic feature selection in a fuzzy rule-based classification 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 Modelling 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 classification with feature selector 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 labelled 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. Modelling 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, FUZZIEEE 2007 (London, UK, 2007).
20. SHEN, Q., AND CHOUCHOULAS, A. A rough-fuzzy approach for generating classification 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 measures 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 neighbour classifiers for hyper spectral satellite imagery. Pattern Recognition Letters, 23 (2002), 183–190.
30. Zhang, Y., Wu, X.-B., Xiang, Z.-R., Hu, W.-L. Design of high-dimensional fuzzy classification systems based on multi-ob jective evolutionary algorithm. Journal of System Simulation, 19 (1), pp. 210-215. 2007.
31. Qinghua Hu, Jinfu Liu and Daren Yu, Mixed feature selection based on granulation and approximation, Knowl. Based Syst., doi:10.1016/j.knosys.2007.07.001. 2007.
32. Qinghua Hu, Zongxia Xie, Daren Yu. Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation. Pattern Recognition 40, pp 3509 – 3521.2007.
33. Luciano Sánchez, M. Rosario Suárez, J. R. Villar and Inés Couso. Some Results about Mutual Information-based Feature Selection and Fuzzy Discretization of Vague Data. IEEE International Conference on Fuzzy Systems. London (UK). 2007.
34. Javier Grande, M. Rosario Suárez, José Ramón Villar. A Feature Selection method using a fuzzy mutual information measure. Advances in Soft Computing 44, pp. 56-63. 2007.
35. Luciano Sánchez, M. Rosario Suárez, José Ramón Villar and Inés Couso. Mutual Information-based Feature Selection and Fuzzy Discretization of Vague Data. Submitted for evaluation to the International Journal of Approximate Reasoning.

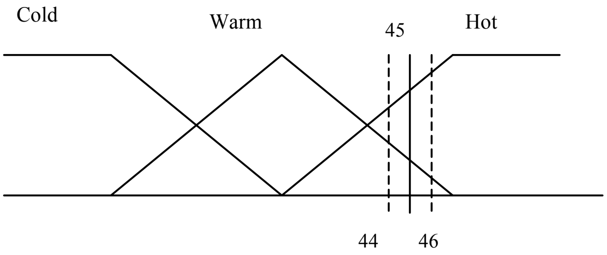

Figure 1 *Relevance of a partitioning scheme in a fuzzy data*

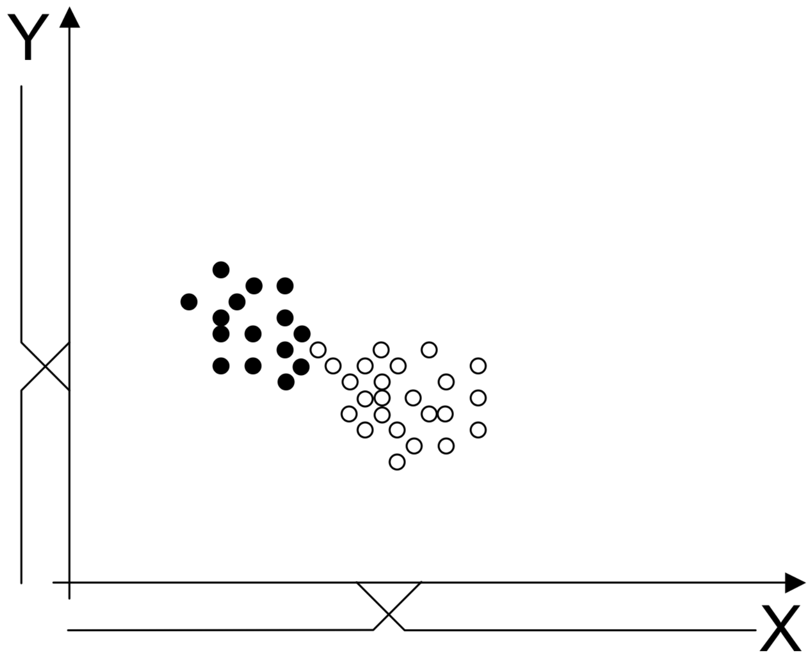

Figure 2 *Two variables two-class problem, with a partitioning schema*

```
Begin
    $F=\{f_1,f_2,\ldots,f_n\}$
    $S=\{\Phi\}$
    Let $\beta$ predefined constant
    For each feature f in F
        calculate MI(Class,f)
    EndFor
    Select the feature f with maximum MI(Class, f)
    Delete $f_i$ from F
    Insert $f_i$ in S
    While |S|<=k
        Calculate the MI(f, s) $\forall$f in F, s in S
        Select feature f that maximizes information
            gain between fi and the class
        Delete f from F
        Insert f in S
    EndWhile
    The output is S
End
```

Figure 3 *The MIFS algorithm*

```
Begin
    $F=\{f_1,f_2,\ldots,f_n\}$
    $S=\{\Phi\}$
    Let $\beta$ predefined constant
    For each feature f in F
        calculate FMI(Class,f)
    EndFor
    Sort  F with respect FMI
    Select a non dominated f in F
    Delete f from F
    Insert f in S
    While  $|S|<=k$
        Calculate the FMI(f, s) $\forall$f in F, s in S
        Calculate the information gain (see text)
        Sort F with respect to the information gain
        Select a non dominated feature  f  in F
        Delete f from F
        Insert f in S
    EndWhile
    The output is S
End
```

Figure 4 *The FMIFS algorithm. Please, refer to the text for the measure information gain.*

	RELIEF	SSGA	MIFS	FMIFS
HEU1	0.295	0.265	0.280	**0.255**
HEU2	0.285	**0.255**	0.265	**0.255**
HEU3	0.275	**0.250**	0.265	0.255
HEU4	0.275	**0.255**	0.265	**0.255**
HEU5	0.275	**0.255**	0.265	**0.255**
REWP	0.280	**0.250**	0.265	0.260
ANAL	0.275	0.260	0.260	**0.245**
GENS	0.270	0.255	0.265	**0.250**
MICH	**0.295**	**0.295**	**0.295**	0.305
PITT	0.285	**0.275**	**0.275**	**0.275**
HYBR	0.295	**0.255**	0.285	**0.255**
ADAB	0.290	**0.260**	0.265	0.265
LOGI	0.260	0.255	**0.250**	0.270
best	1	9	3	8

Figure 5 *The average classification error after the 10 k fold cross validation of* ***GERMAN*** *rule-based classifiers after performing a feature selection, with the original MIFS algorithm and with the modified version proposed in this paper.*

	RELIEF	SSGA	MIFS	FMIFS
HEU1	0.500	**0.176**	0.323	**0.176**
HEU2	0.411	0.176	0.323	**0.147**
HEU3	0.235	0.147	0.264	**0.117**
HEU4	0.205	0.235	0.205	**0.176**
HEU5	0.176	**0.147**	0.176	**0.147**
REWP	0.088	**0.058**	0.117	**0.058**
ANAL	0.235	**0.088**	0.235	0.147
GENS	**0.029**	0.147	0.176	0.117
MICH	0.647	**0.147**	0.617	0.176
PITT	0.205	0.058	0.058	**0.029**
HYBR	**0.029**	**0.029**	0.176	0.088
ADAB	0.058	**0.000**	0.058	0.058
LOGI	0.058	**0.029**	0.058	0.058
best	2	8	0	7

Figure 6 *The average classification error after the 10 k fold cross validation of* ***WINE*** *rule-based classifiers after performing a feature selection, with the original MIFS algorithm and with the modified version proposed in this paper.*

	RELIEF	SSGA	MIFS	FMIFS
HEU1	0.328	0.200	0.200	**0.185**
HEU2	0.314	0.185	0.200	**0.142**
HEU3	0.285	0.157	0.200	**0.128**
HEU4	0.285	0.157	0.200	**0.128**
HEU5	0.285	0.157	0.200	**0.128**
REWP	0.200	0.142	0.185	**0.128**
ANAL	0.257	**0.157**	0.185	0.171
GENS	0.157	0.128	0.185	**0.100**
MICH	0.428	0.328	0.357	**0.200**
PITT	0.228	**0.114**	0.157	**0.114**
HYBR	0.214	**0.114**	0.142	0.128
ADAB	**0.114**	0.514	0.514	0.514
LOGI	0.142	0.100	0.171	**0.085**
best	1	3	0	10

Figure 7 *The average classification error after the 10 k fold cross validation of* ***ION*** *rule-based classifiers after performing a feature selection, with the original MIFS algorithm and with the modified version proposed in this paper.*

	RELIEF	SSGA	MIFS	FMIFS
HEU1	0.289	0.302	**0.276**	0.302
HEU2	0.289	0.289	**0.276**	0.289
HEU3	0.276	**0.263**	0.276	**0.263**
HEU4	0.276	**0.263**	0.276	**0.263**
HEU5	0.276	**0.263**	0.276	**0.263**
REWP	0.269	**0.263**	0.276	**0.263**
ANAL	0.269	**0.263**	0.276	**0.263**
GENS	0.263	0.263	0.269	**0.243**
MICH	**0.355**	**0.355**	**0.355**	**0.355**
PITT	**0.230**	0.243	0.256	0.250
HYBR	0.256	**0.243**	0.276	0.276
ADAB	0.243	**0.217**	0.223	**0.217**
LOGI	0.250	**0.217**	0.243	**0.217**
best	2	9	3	9

Figure 8 *The average classification error after the 10 k fold cross validation of* ***PIMA*** *rule-based classifiers after performing a feature selection, with the original MIFS algorithm and with the modified version proposed in this paper.*

	RELIEF	SSGA	MIFS	FMIFS
HEU1	0.300	0.300	0.350	**0.225**
HEU2	0.275	0.325	0.325	**0.200**
HEU3	0.250	0.250	0.300	**0.175**
HEU4	0.250	0.250	0.300	**0.175**
HEU5	0.250	0.250	0.300	**0.175**
REWP	0.275	0.300	0.350	**0.200**
ANAL	0.375	0.325	0.350	**0.225**
GENS	0.300	0.250	0.250	**0.175**
MICH	**0.300**	**0.300**	0.350	**0.300**
PITT	**0.275**	0.300	0.325	**0.275**
HYBR	0.325	0.250	0.350	**0.225**
ADAB	0.300	0.250	0.350	**0.150**
LOGI	0.250	0.250	0.325	**0.200**
best	2	1	0	13

Figure 9 *The average classification error after the 10 k fold cross validation of* ***SONAR*** *rule-based classifiers after performing a feature selection, with the original MIFS algorithm and with the modified version proposed in this paper.*