An analysis on the use of pre-processing methods in evolutionary fuzzy systems for subgroup discovery

Cristóbal J. Carmona [a,*], Julián Luengo [b], Pedro González [a], María José Del Jesus [a]

[a] Department of Computer Science, University of Jaen, Campus las Lagunillas, 23071 Jaen, Spain
[b] Department of Civil Engineering, University of Burgos, 09006 Burgos, Spain

ARTICLE INFO

Keywords:
Subgroup discovery
Evolutionary fuzzy systems
Missing values

ABSTRACT

Subgroup discovery is a descriptive data mining technique which aims at obtaining interesting rules through supervised learning. In general, there are no works analysing the consequences of the presence of missing values in data in this task, although improper handling of this type of data in the analysis may introduce bias and can result in misleading conclusions being drawn from a research study. This paper presents a study on the effect of using the most relevant approaches for pre-processing of missing values in a determined group of algorithms, the evolutionary fuzzy systems for subgroup discovery.

The experimental study presented in this paper show that, among the methods studied, the KNNI pre-processing approach for missing values obtains the best results in evolutionary fuzzy systems for subgroup discovery.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Subgroup discovery (SD) (Kloesgen, 1996; Wrobel, 1997) is a descriptive rule induction technique using supervised learning. The main purpose is to extract descriptive knowledge with respect to a property of interest, i.e. this technique attempts to obtain descriptive subgroups which are statistically most interesting (since they differ from the complete data set) with respect to a target variable or class. SD algorithms are widely employed throughout the literature, due to their capability to describe problems with a high interpretability respect to a property of interest. In addition, these algorithms are well known for being able to deal with real-world applications like bioinformatics, marketing, or e-learning problems, among others. A complete review of these applications can also be observed in Herrera, Carmona, González, and Del Jesus (2011).

Evolutionary fuzzy systems (EFSs) have been successfully applied to the SD task. An EFS (Herrera, 2008) uses basically evolutionary algorithms (Eiben & Smith, 2003) for learning or tuning fuzzy systems. EFSs are appropriated for SD because evolutionary algorithms handle appropriately the relations between variables, and the use of fuzzy logic by means of descriptive fuzzy rules allows the representation of knowledge in a similar way to human reasoning, leading to the obtaining of more interpretable and actionable solutions in SD.

The quality of data handled greatly influences the results of the subgroup discovery algorithms. One of the most important aspects regarding the quality of data is the presence of missing values (MVs). However, it is not usual to take into account the presence of MVs in data for SD techniques and few studies have been carried out considering this problem (Atzmueller & Puppe, 2006). The inappropriate handling of missing data in the analysis may introduce bias and can result in misleading conclusions being drawn from a research study, and can also limit the generalisability of the research findings (Wang & Wang, 2010). In this way Di-Nuovo (2011) presents an analysis with fuzzy C-means in missing data for a psychological problem.

The problem of the MVs is very important when using EFSs, since most fuzzy systems cannot work directly with incomplete data sets. In addition, for real-world problems it is possible that no valid (complete) cases would be present in the data set (García-Laencina, Sancho-Gómez, & Figueiras-Vidal, 2009). It has been shown that the existence of incomplete data in both the training set as the test set (as in both) affect the prediction accuracy of learned classifiers (Gheyas & Smith, 2010) and this also applies to SD. The severity of this problem depends in part on the proportion of missing data. Real problems with missing data can be observed for example in Doh-Soon and Kwang-Jae (2012) or Yongsong, Shichao, Xiaofeng, Jilian, and Chengqi (2009).

Therefore it follows that the incomplete data must be treated previously, in order to avoid these drawbacks. It will also be necessary a pre-processing stage in which data is prepared and cleaned (Pyle, 1999), in order to be not only useful but clear enough for the knowledge extraction process.

* Corresponding author. Tel.: +34 953 211956; fax: +34 953 212472.
  E-mail address: ccarmona@ujaen.es (C.J. Carmona).

The simplest way of dealing with missing values is to discard the examples that contain them. However, this method is only practical when data contains a relatively small number of examples with MVs and when analysis of complete examples does not lead to a significant bias during the inference (Little & Rubin, 1987).

Two additional approaches are usually found in the literature to deal with MVs (Farhangfar, Kurgan, & Pedrycz, 2007): The use of maximum likelihood procedures to estimate the parameters of the model based on complete data, and the imputation of MVs which aims to fill in the MVs with estimated values. The operation of this second type of approach is based on that, in many cases, the attributes of a data set are not independent from each other, so that the identification of relationships among attributes allows to determine appropriate values for MVs.

The effect of using a large set of imputation methods on SD techniques based on EFSs is studied in order to determine the best imputation procedure for each one. In this paper, SD algorithms based on EFSs presented throughout the literature are employed: SDIGA (Del Jesus, González, Herrera, & Mesonero, 2007c), MESDIF (Del Jesus, González, & Herrera, 2007b) and NMEEF-SD (Carmona, González, Del Jesus, & Herrera, 2010). In order to perform this analysis, we use a large bunch of data sets, 19 in total. First, we analyse the use of the different imputation strategies versus case deletion and the total lack of missing data treatment, for a total of 14 imputation methods. Therefore, each SD technique is applied over the results of the 14 imputation methods. All the imputation and SD algorithms used are publicly available in the KEEL (Alcalá-Fdez et al., 2009; Alcalá-Fdez et al., 2011) software.[1]

The rest of the paper is organised as follows: Section 2 introduces the SD task, its main properties and quality measures and the EFSs algorithms for SD analysed in this paper. Section 3 presents MVs and different pre-processing approaches to optimize data sets with this type of values. In Section 4 the experimental study is presented. Finally, some concluding remarks are outlined in Section 5.

2. Subgroup discovery

SD is a descriptive technique of data mining where supervised learning is employed in order to find interesting relations with respect to a property of interest for the experts. SD falls within the recently introduced concept of "supervised descriptive rule discovery" (Kralj-Novak, Lavrac, & Webb, 2009) along with contrast set mining (Bay & Pazzani, 2001) and emerging pattern mining (Dong & Li, 2005).

This section presents the definition and main properties of SD and furthermore, as this paper is focused on the use of EFSs for SD, existing EFSs algorithms for SD, their knowledge representation and the quality measures used are presented.

2.1. Definition of subgroup discovery

The concept of SD was initially introduced by Kloesgen (1996) and Wrobel (1997), and more formally defined by Siebes (1995) but using the name data surveying for the discovery of interesting subgroups. It can be defined as (Wrobel, 2001):

"In subgroup discovery, we assume we are given a so-called population of individuals (objects, customer, ...) and a property of those individuals we are interested in. The task of subgroup discovery is then to discover the subgroups of the population that are statistically "most interesting", i.e. are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest."

Considering this definition, the main property of this task is the search of partial relations where the majority of the examples corresponding to the property of interest or target variable will be covered. In addition, these relations should be interesting and have an unusual behavior.

In order to represent the knowledge, SD employs a rule (R) which consists of an induced subgroup description. It can be formally defined as:

$R : Cond \rightarrow TargetVar$

where *TargetVar* is a value for the variable of interest (target variable) for the SD task and *Cond* is commonly a conjunction of features (attribute-value pairs) capable of describing an unusual statistical distribution with respect to the *TargetVar*.
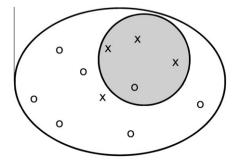
As an example of subgroup, a rule is represented in Fig. 1, where two values for the target variable are possible (*TargetVar* = x and *TargetVar* = o). In this representation a subgroup for the first value of the target variable are can be observed, where the rule attempts to cover a high number of objects with a single function (described in Fig. 1 as a circle). As can be observed the subgroup does not cover all the examples for the target value x besides some examples covered do not correspond to the target value, but the form of this function is uniform and simple. Furthermore, the true positive rate corresponding to the value of the target variable is high, with a value of 75%.

The main elements of a SD approach were defined by Atzmueller et al. in Atzmueller, Puppe, and Buscher (2004). The study and configuration of these elements is very important in order to develop a new approach for SD task. They are as follows:

- Type of the target variable, as for example binary, categorical, numerical and so on.
- Description language, where the most representative is the use of pairs attribute-value because rules must be simple.
- Search strategy, a very important aspect since the search space grows exponentially with the number of possible expressions that can be part of a subgroup description. In this paper we focus on evolutionary algorithms.
- Quality measures. It is the most important element because they guide the search process and evaluate the quality of the subgroups obtained. In Herrera et al. (2011) is presented a complete study respect to the quality measures presented throughout the literature.

2.2. Evolutionary fuzzy systems in subgroup discovery

Evolutionary algorithms are stochastic algorithms for optimizing and search which are based on the natural evolution process. These algorithms were introduced by Holland (1975). The hybridisation between fuzzy logic (Zadeh, 1994) and evolu-

Fig. 1. Representation of a subgroup discovery rule with respect to a value (x) of the target variable.

tionary algorithms is known as EFS (Herrera, 2008). Evolutionary fuzzy systems are fuzzy systems enhanced by a learning process based on an evolutionary algorithm. This type of algorithms provide novel and useful tools for pattern analysis and for extracting new kinds of useful information.

Throughout the literature different EFSs for the SD task have been presented, which are described below:

- SDIGA (Del Jesus et al., 2007c) is an evolutionary fuzzy rule induction system. The quality measures used for the SD task are adaptations of the ones used in algorithms for association rule induction, confidence and support, and can also use other measures such as interest, significance, sensitivity or unusualness. The algorithm evaluates the quality of the rules by means of a weighted average of the measures selected. An analysis of different combinations of quality measures can be observed in Carmona, González, Del Jesus, and Herrera (2009). SDIGA uses linguistic rules (Del Jesus, González, & Herrera, 2007a) as description language to specify the subgroups.
- MESDIF (Berlanga, Del Jesus, González, Herrera, & Mesonero, 2006; Del Jesus et al., 2007b) is a multi-objective genetic algorithm for the extraction of fuzzy rules describing subgroups. The algorithm extracts a variable number of different rules expressing information on a single value of the target variable. The search is based on the multi-objective SPEA2 (Zitzler, Laumanns, & Thiele, 2002) approach, and so applies the concepts of elitism in the rule selection (using a secondary or elite population) and the search for optimal solutions in the Pareto front. It can use several quality measures at a time to evaluate the rules obtained, like confidence, support, sensitivity, significance or unusualness.
- NMEEF-SD (Carmona, González, Del Jesus, & Herrera, 2009; Carmona et al., 2010) is an evolutionary fuzzy system whose objective is to extract descriptive fuzzy and/or crisp rules for the SD task, depending on the type of variables present in the problem. NMEEF-SD has a multi-objective approach whose search strategy is based on NSGA-II (Deb, Pratap, Agrawal, & Meyarivan, 2002), which uses a non-dominated sorting approach and elitism. This algorithm uses specific operators to promote the extraction of simple, interpretable and high quality rules. It allows a number of quality measures to be used both for the selection and the evaluation of rules within the evolutionary process, including confidence, support, sensitivity, significance, and unusualness.

Table 1 presents a comparative analysis among the most important elements used by the EFSs approaches for SD. These algorithms are implemented in the software tool KEEL.

Moreover, these algorithms have been applied in several real-world problems like e-learning (Carmona, González, Del Jesus, & Ventura, 2011; Romero, González, Ventura, Del Jesus, & Herrera, 2009) or medicine (Carmona, González, Del Jesus, Navío, & Jiménez, 2011), for example.

2.3. Description language for evolutionary fuzzy systems in subgroup discovery

EFSs developed so far use pairs of attribute-value as description language, where the values are represented through fuzzy logic (Zadeh, 1994) for continuous attributes.

In this way, this type of algorithms obtain fuzzy rules (Hüllermeier, 2005) by means of linguistic variables allowing the use of numerical features in data mining processes. Continuous variables are considered linguistic, and the fuzzy sets corresponding to the linguistic labels can be specified by the user or defined by means of a uniform partition, if the expert knowledge is not available. Fig. 2 shows an example using uniform partitions with triangular membership functions for a variable with three linguistic labels: *Low*, *Medium* and *High*. Moreover, it is important to remark that if the data set is formed only by discrete attributes, crisp rules are obtained.

Below equation shows a fuzzy subgroup with different variables and linguistic labels associated for each one. The notation used in the fuzzy rule expression is explained below:

$$R : IF\ X_1 = (LL_1^1)\ AND\ X_6 = (LL_6^2)\ THEN\ Target_{value}$$

- $X = \{X_m/m = 1, \ldots, n_v\}$ is a set of features used to describe the subgroups, where n_v is the number of properties or variables. These variables can be categorical or numerical.
- $T = \{Target_{value}/j = 1, \ldots, n_c\}$ is a set of values for the target variable, where n_c is the number of values.
- $E = \left\{ E^k = \left(e_1^k, e_2^k, \ldots, e_{n_v}^k, Target^k \right) /k = 1, \ldots, n_s, Target^k \in T \right\}$ is a set of examples, where $Target^k$ is the value of the target variable for the example E^k (i.e. the class for this example) and n_s is the number of examples.
- $LL_{n_v}^{l_{n_v}}$ is the linguistic label number l_{n_v} of the variable n_v.

As can be observed the above equation (Rule) shows a fuzzy subgroup with two variables, where the first variable (X_1) takes as value its first linguistic label $\left(LL_1^1 \right)$, and the second variable (X_6) takes its second linguistic label $\left(LL_6^2 \right)$.

2.4. Quality measures for evolutionary fuzzy systems in subgroup discovery

As we have mentioned previously, the most important element for the obtaining of good results in a SD task is the set of quality measures used because they guide the search process and evaluate the quality of the subgroups obtained.

Table 1
Comparison of properties for evolutionary fuzzy systems in subgroup discovery task.

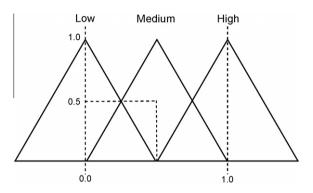| Feature	SDIGA	MESDIF	NMEEF-SD
Evolutionary model	IRL approach	MOEA based on SPEA2 model	MOEA based on NSGA-II model
Knowledge representation	Canonical or DNF fuzzy rules	Canonical or DNF fuzzy rules	Canonical or DNF fuzzy rules
Individual coding	Chromosome = rule	Chromosome = rule	Chromosome = rule
Quality measures used as objectives	Aggregation function with different measures among: Coverage, significance, unusualness, accuracy, sensibility, crisp support, fuzzy support, crisp confidence and fuzzy confidence	Selected by the user among: Coverage, significance, unusualness, accuracy, sensibility, crisp support, fuzzy support, crisp confidence and fuzzy confidence	Selected by the user among: Coverage, significance, unusualness, accuracy, sensibility, crisp support, fuzzy support, crisp confidence and fuzzy confidence
Selection	Steady step GA	Tournament	Tournament
Operators	Biased mutation, two point crossover and local search	Biased mutation and two point crossover	Biased mutation, two point crossover and re-initialisation based on coverage

Fig. 2. Example of fuzzy partition for a continuous variable with three labels.

As specified by the authors, each algorithm uses some quality measures to guide the search process. However, in order to measure the quality of the results of the SD algorithms considered in this paper, the following quality measures are employed:

- *Number of rules* (n_r), which measures the number of induced rules.
- *Number of variables* (n_v). This quality measure considers the average number of variables in the antecedent of the rules, computed as the average number of variables for each rule of the set of rules.
- *Significance*. This measure indicates how significant is a finding, and it is measured by the likelihood ratio of a rule (Kloesgen, 1996).

$$Sign(R) = 2 \cdot \sum_{k=1}^{n_c} n(TargetVar_k \cdot Cond)$$
$$\cdot \log \frac{n(TargetVar_k \cdot Cond)}{n(TargetVar_k) \cdot p(Cond)} \qquad (1)$$

where $n(TargetVar \cdot Cond)$ is the number of examples satisfying the conditions and also belonging to the value for the target variable in the rule, $n(TargetVar)$ is the number of examples of the target variable, $p(Cond)$ is used as a normalized factor, and n_c is the number of values of the target variable. It must be noted that although each rule corresponds to a specific *TargetVar*, the significance measures the novelty in the distribution impartially, for all the values.

- *Unusualness*: This measure is defined as the weighted relative accuracy of a rule (Lavrac, Flach, & Zupan, 1999). It can be computed as:

$$Unus(R) = \frac{n(Cond)}{n_s} \cdot \left(\frac{n(TargetVar \cdot Cond)}{n(Cond)} - \frac{n(TargetVar)}{n_s} \right) \qquad (2)$$

The unusualness of a rule can be described as the balance between the coverage of the rule $p(Cond_i)$ and its accuracy gain $p(TargetVar \cdot Cond) - p(TargetVar)$.

- *Sensitivity*: This measure is the proportion of actual matches been correctly classified (Kloesgen, 1996). It can be computed as:

$$Sens(R) = TPr = \frac{TP}{Pos} = \frac{n(TargetVar \cdot Cond)}{n(TargetVar)} \qquad (3)$$

This quality measure was used in Del Jesus et al. (2007c) as *Support based on the examples of the class*. Sensitivity combines precision and generality related to the target variable.

- *Fuzzy Confidence*: It measures the relative frequency of examples satisfying the complete rule among those satisfying only

the antecedent in fuzzy rules (Del Jesus et al., 2007c). This can be computed as:

$$FCnf(R) = \frac{\sum_{E^k \in E/E^k \in TargetVar} APC(E^k, R)}{\sum_{E^k \in E} APC(E^k, R)} \qquad (4)$$

The following assumptions are important to understand this fuzzy quality measure:

- An example E^k verifies the APC of a rule R_i if

$$APC(E^k, R_i) = T\left(\mu_{LL_1^1}(e_1^k), \ldots, \mu_{LL_{n_v}^{l_{n_v}}}(e_{n_v}^k) \right) > 0 \qquad (5)$$

where *APC* (Antecedent Part Compatibility) is the degree of compatibility between an example and the antecedent part of a fuzzy rule, i.e. the degree of membership for the example to the fuzzy subspace delimited by the antecedent part of the rule, where:

  * $\mu_{LL_{n_v}^{l_{n_v}}}(e_{n_v}^k)$ is the degree of membership for the value of the feature n_v for example E^k to the fuzzy set corresponding to the linguistic label l_{n_v} for this variable (n_v);
  * T is the *t-norm* selected to represent the meaning of the AND operator (the fuzzy intersection) in our case the minimum *t-norm*.
- An example E^k is covered by a rule R_i if

$$APC(E^k, R_i) > 0 \; AND \; E^k \in Class_j. \qquad (6)$$

This means that an example is covered by a rule if the example has a degree of membership higher than 0 to the fuzzy input subspace delimited by the antecedent part of the fuzzy rule, and the value indicated in the consequent part of the rule agrees with the value of the target feature for the example. For the categorical variables, the degrees of membership are 0 or 1.

Quality measures described above have been selected for comparison in this study based on three important guidelines:

1. Interpretability. This guideline measures the number of rules and variables obtained by a SD model. Optimal interpretability for a SD approach is the obtention of few rules containing a low number of variables. This helps the experts to understand and use the knowledge extracted because algorithms search for simple and interpretable subgroups through partial relations.
2. Relation between sensitivity and confidence. This objective quantifies a good compromise between both measures, i.e. the algorithm must achieve the best possible relation between sensibility and confidence. Both quality measures are crucial to give the experts subgroups covering the largest number of correctly described examples. This balance is difficult to obtain as it is usual that the value for a measure is deteriorated when the other is improved.
3. Interest or novelty. This final guidelines is related to the search of interesting and unusualness relations in data. A SD model must contribute novel knowledge to the problem. Fundamentally, this objective could be measured with a wide number of quality measures as novelty, interest or significance, among others. Nevertheless, it is important to highlight the use of the unusualness to measure this objective because it also contributes with generality and confidence to the problem.

3. Missing values

MVs are one of the most influential factors in data quality when data mining processes need to be applied. An inappropriate

handling of missing data may introduce bias in the analysis and the obtaining of misleading conclusions. In this section we set the basis of our study in accordance with the MV literature: The fundamental aspects in the MVs treatment based on the MV introduction mechanisms are indicated, the influence of MVs in subgroup discovery algoritms based on EFSs is described and a brief description of the imputation methods for MVs used in our study is introduced.

3.1. Treatment of missing values

A missing data may represent an unknown value (as a result of the data is left empty), a data that has not yet been assigned a value, an error in the conversion process of the data, or an undefined value resulting from the operation on other missing values or produced by prohibited operations. In order to understand the phenomenon and select the most appropriated method to handle MVs, it is important to categorize the mechanisms which lead to the introduction of MVs (Little & Rubin, 1987). As Little and Rubin (1987) stated, there are three different mechanisms for missing data induction:

1. Missing completely at random (MCAR), when the distribution of an example having a missing value for an attribute does not depend on either the observed data or the missing data. In this case, examples with complete data are indistinguishable from cases with incomplete data.
2. Missing at random (MAR), when the distribution of an example having a missing value for an attribute depends on the observed data, but does not depend on the missing data. In this case, examples with incomplete data differ from cases with complete data, but missing data are predictable from other variables in the database.
3. Not missing at random (NMAR), when the distribution of an example having a missing value for an attribute depends on the missing values. In this case, the pattern producing missing values is nonrandom and it is nos possible to predict the values from other variables in the database.

In the case of the MCAR mode, the assumption is that the underlying distributions of missing and complete data are the same, while for the MAR mode they are different, and the missing data can be predicted by using the complete data (Little & Rubin, 1987).

The treatment of missing data can be handled in three different ways (Li, Deogun, Spaulding, & Shuart, 2004):

- The simplest approach is to discard the examples with missing data in their attributes. Therefore deleting attributes with high levels of missing data is enclosed in this category too. This is only practical when data contains few MVs and analysing only complete examples does not lead to a significant bias during the inference.
- A second approach is the use of maximum likelihood procedures to estimate the parameters of the model based on complete data and then been used to impute by means of sampling.
- Finally, the imputation of MVs, whose objective is to estimate the values to fill in the MVs. To do so, this method takes advantage of the fact that, in most cases, the attributes of a data set are not independent from each other, so that the identification of relationships among attributes allows to determine appropriate values for MVs. This is the most used approach.

As Farhangfar, Kurgan, and Dy (2008) and Matsubara, Prati, Batista, and Monard (2008) state, it is only in the MCAR mechanism case where the analysis of the remaining complete data (ignoring incomplete data) could give a valid inference due to the assump-tion of equal distributions. That is, case and attribute removal with missing data should be applied only if the missing data is MCAR, as both of the other mechanisms could potentially lead to information loss that would lead to the generation of a biased/incorrect model (i.e. a model based on a different distribution).

Another solution is to convert the missing values to a new value (encode them into a new numerical value), but such a simplistic method was shown to lead to serious inference problems (Schafer, 1997). On the other hand, if a significant number of examples contain missing values for a relatively small number of attributes, it may be beneficial to perform imputation (filling-in) of the missing values. In order to do so, the assumption of MAR randomness is needed, as Little and Rubin (1987) observed in their analysis.

Throughout the literature different approaches can be found in order to handle missing values, for example in Hai and Shouhong (2009) a proposal based on a rough set rule induction is presented, in Yongsong, Shichao, Xiaofeng, Jilian, and Chengqi (2009) a kernel-based missing data imputation can be observed, or in Yan, Yiyuan, Shuxue, Shipin, and Yifen (2011) an approach based on bayesian network with missing data is shown.

3.2. Effect of missing values in subgroup discovery algorithms

As we have previously mentioned, the presence of MVs harms the capabilities of the model obtained when applying data mining techniques. It has been shown that MVs reduce the accuracy of predictive models (Batista & Monard, 2003a; Gheyas & Smith, 2010). In addition, the completeness of the obtained knowledge is also reduced (Pyle, 1999) if the MVs are not correctly treated. Three types of problems are usually associated with MV in data mining (Barnard & Meng, 1999): (1) loss of efficiency; (2) complications in handling and analysing the data; and (3) bias resulting from differences between missing and complete data.

This is specially critical in a descriptive task like SD, where the completeness of the information presented is crucial, as well as its generalization abilities. Therefore it follows that incomplete data must be previously treated in order to avoid these drawbacks. However, up to this study the presence of MVs in the data has been slightly considered when tackling the SD task. We can find a unique precedent in the SD-Map algorithm (Atzmueller & Puppe, 2006), which is capable of handling MVs during the learning phase. The rest of the proposals of SD algorithms are not able to cope with the presence MVs, disabling their use in many real world data sets.

Among the different approaches available to handle MVs, special attention of this paper is focused on imputation methods. A fundamental advantage of this approach is that the missing data treatment is independent of the learning algorithm used. This allows the user to select the most appropriate method for each situation he faces.

SD algorithms employed in this paper use fuzzy systems to improve the robustness of the process, as they are capable of managing certain degree of imperfections in the data. However, in spite of the fuzzy systems being imperfect data tolerant, it has been proved that the use of imputation techniques is necessary in order to achieve the best results (Luengo, Sáez, & Herrera, 2012). Using imputation techniques avoids the necessity of modifying the SD algorithm which will be applied afterwards, thus providing two advantages:

1. Data needs to be pre-processed only once. Thereafter, pre-processed data can be used by every SD algorithm needed.
2. Avoids the modification of each SD algorithm to handle MVs, which is not possible in some cases due to the definition of the method.

3.3. Description of the imputation methods used

There is a wide family of imputation methods, from simple imputation techniques like mean substitution, *K*-Nearest Neighbor, etc., to those which analyse the relationships between attributes such as support vector machines-based, clustering-based, logistic regressions, maximum-likelihood procedures and multiple imputation (Batista & Monard, 2003a; Farhangfar et al., 2008). Trying a vast amount of them in order to decide the best approach would be very time consuming. Previous studies have shown that particular learning algorithms usually obtain more benefit from a small amount of imputation techniques. Therefore, establishing which imputation techniques are the more appropriate for the SD methods is needed.

In our case we will use single imputation methods, due to the time complexity of the multiple imputation schemes, and the assumptions they make regarding data distribution and MV randomness; that is, that we should know the underlying distributions of the complete data and missing data prior to their application.

In the following are briefly described the imputation methods considered in this study:

- Do Not Impute (DNI). As its name indicates, all the missing data remains unreplaced, so the algorithms must use their default MVs strategies. The objective is to verify whether imputation methods allow the algorithms to perform better than when using the original data sets. As a guideline, in Grzymala-Busse and Hu (2000) a previous study of imputation methods is presented.
- Case deletion or Ignore Missing (IM). Using this method, all instances with at least one MV are discarded from the data set.
- Global most common attribute value for symbolic attributes, and Global average value for numerical attributes (MC) (Grzymala-Busse, Goodwin, Grzymala-Busse, & Zheng, 2005). This method is very simple: For nominal attributes, the MV is replaced with the most common attribute value, and numerical values are replaced with the average of all values of the corresponding attribute.
- Concept most common attribute value for symbolic attributes, and Concept average value for numerical attributes (CMC) (Grzymala-Busse et al., 2005). As stated in the *MC* method, the MV is replaced by the most repeated value for nominal attributes, or by the mean value for numerical ones, but considering only the instances with the same class as the reference instance.
- Imputation with fuzzy *K*-means Clustering (FKMI) (Acuna & Rodríguez, 2004; Li et al., 2004). In fuzzy clustering, each data object has a membership function which describes the degree to which this data object belongs to a certain cluster. In the process of updating membership functions and centroids, FKMI's only take into account complete attributes. In this process, the data object cannot be assigned to a concrete cluster represented by a cluster centroid (as is done in the basic *K*-mean clustering algorithm), because each data object belongs to all *K* clusters with different membership degrees. FKMI replaces non-reference attributes for each incomplete data object based on the information about membership degrees and the values of cluster centroids.
- Imputation with *K*-Nearest Neighbor (KNNI) (Batista & Monard, 2003b). Using this instance-based algorithm, every time an MV is found in a current instance, KNNI computes the *k* nearest neighbors and a value from them is imputed. For nominal values, the most common value among all neighbors is taken, and for numerical values the average value is used. Therefore, a proximity measure between instances is needed for it to be defined. The euclidean distance (it is a case of a L_p norm distance) is the most commonly used in the literature.

An extensive and detailed description of these methods can be found on the web page <http://sci2s.ugr.es/MVDM>, and a PDF file with the original source paper descriptions is present on the web page formerly named "Imputation of Missing Values. Methods' Description". A more exhaustive bibliography section is also available on the mentioned web page.

4. Experimental study

This section presents a complete experimental study with the different EFSs for SD presented throughout the literature. Several data sets, both with MVs and with induced MVs are used. As the results obtained by the SD algorithms will depend on the previous pre-processing step of the data sets, the objective is to determine which of the imputation methods for the treatment of data sets with MVs is more appropriated as a pre-processing step in the SD task using EFSs. To do so, the imputation approaches presented in the previous section are applied to the selected data sets with MVs, and then the SD algorithms are used.

Therefore, this experimental study is divided in two sections. First, the experimental framework and the parameters used for the different algorithms and data sets is indicated, and then the results obtained by the different EFSs for the SD task are shown.

4.1. Experimental framework

The experimentation was undertaken with real data sets from the KEEL (<http://www.keel.es>) repository.

The properties of the data sets are presented in Table 2 including: Number of variables (n_v), number of discrete variables (n_{vD}), number of continuous variables (n_{vC}), number of classes (n_c), number of examples (n_s) and percentage of examples with missing values (%MV) which indicates percentage of instances with at least one MV. Among the complete data sets analysed, *Iris*, *Pima*, *New-Thyroid*, *Ecoli*, *German*, *Magic* and *Shuttle* have missing values induced. These data sets are modified versions from the original ones, where a 10% of values have been randomly removed (only training partitions present missing values, while test partitions remain unchanged). With respect to the remaining data sets, containing MVs, there is no information available about the generation mechanism of the MVs.

In order to perform the experimental study, SD algorithms are executed using a ten fold cross-validation (10-fcv) procedure.

Table 2
Properties of the data sets used from the KEEL repository.

Name	n_v	n_{vD}	n_{vC}	n_c	n_s	%MV
Adult	14	8	6	2	48842	7.41
Bands	19	6	13	2	539	32.28
Breast	9	9	0	2	286	3.15
Cleveland	13	0	13	5	303	1.98
Crx	15	12	3	2	690	5.36
Dermatology	34	34	0	6	366	2.19
Ecoli	7	0	7	8	336	48.21
German	20	20	0	2	1000	80.00
Hepatitis	19	13	6	2	155	48.39
Horse-colic	23	16	7	2	368	98.10
Housevotes	16	16	0	2	435	46.67
Iris	4	0	4	3	150	32.67
Magic	10	0	10	2	1902	58.20
Mammographic	5	5	0	2	961	13.63
Mushroom	22	22	0	2	8124	30.53
New-Thyroid	5	1	4	3	215	35.35
Pima	8	0	8	2	768	50.65
Shuttle	9	9	0	7	2175	55.95
Wisconsin	9	9	0	2	699	2.29

Table 3
Parameters of algorithms employed.

Algorithm	Parameters
SDIGA	Population size = 50, evalutions = 10000, crossover probability = 0.60, mutation probability = 0.1, minimum confidence = 0.6, representation of the rule = canonical and linguistic labels = 3, objective1 = sensitivity (weight = 0.4), objective2 = fuzzy confidence (weight = 0.3) and objective3 = unusualness (weight = 0.3).
MESDIF	Population size = 50, evalutions = 10000, crossover probability = 0.60, mutation probability = 0.01, elite population size = 3, representation of the rule = canonical, linguistic labels = 3, objective1 = sensitivity, objective2 = fuzzy confidence and objective3 = unusualness.
NMEEF-SD	Population size = 50, evalutions = 10000, crossover probability = 0.60, mutation probability = 0.1, minimum confidence = 0.6, representation of the rule = canonical, linguistic labels = 3, objective1 = sensitivity and objective2 = unusualness.

Table 4
Average results obtained for SD algorithms with different missing values approaches.

Algorithm	MV_{method}	n_r	n_v	SIGN	UNUS	SENS	FCNF
SDIGA	CMC	4.41	**3.08**	16.199	0.051	0.731	0.561
	DNI	4.34	3.73	**18.297**	**0.056**	0.679	0.548
	FKMI	4.41	3.14	16.266	0.052	0.761	0.570
	IM	4.24	3.56	10.679	0.052	0.683	0.546
	KNNI	**4.16**	3.12	16.508	0.052	**0.760**	**0.576**
	MC	4.40	3.22	15.683	0.048	0.734	0.553
MESDIF	CMC	9.13	6.72	8.869	0.028	0.360	0.477
	DNI	9.16	6.36	9.542	0.030	0.364	0.445
	FKMI	9.16	6.84	9.059	0.026	0.361	0.456
	IM	9.16	6.96	8.876	0.028	0.350	0.439
	KNNI	**9.00**	6.72	9.814	**0.034**	**0.403**	**0.490**
	MC	9.16	6.87	9.574	0.028	0.377	0.455
NMEEF-SD	CMC	10.99	3.62	13.534	**0.109**	0.909	0.780
	DNI	9.78	3.44	13.111	0.103	0.879	0.750
	FKMI	**7.19**	**3.11**	12.718	0.102	0.901	0.764
	IM	14.01	3.59	13.592	0.104	0.869	0.777
	KNNI	9.55	3.50	**13.909**	0.106	**0.912**	**0.785**
	MC	9.82	3.54	13.247	0.104	0.899	0.770

Therefore, the results shown for the experiments are the average of the results obtained for each data set for the different partitions. The results shown for each SD algorithm are the average of 50 executions (5 executions per group of cross validation).

Parameters used in the execution of the SD algorithms, they are presented in Table 3.

Quality measures analysed and presented in the result tables are the average results for the rule sets in the total data sets studied with the following initials: FCNF for fuzzy confidence, SENS for sensitivity, SIGN for significance and UNUS for unusualness.

4.2. Results obtained

The results of the SD algorithms based on EFSs for this experimental study are presented in Table 4. In this table, the name of the SD algorithm (*Algorithm*) used, the pre-processing method for MVs (*MV_method*) employed, and the results of the quality measures explained above are shown. The complete results obtained for each algorithm in the different data sets are available in the Website <http://simidat-web.ujaen.es/MVs/ESWA12>.

A summary of the results obtained for each algorithm is presented in order to perform a comprehensible analysis of this experimental study. This analysis is carried out according to the three guidelines introduced previously, i.e. "interpretability", "interest" and "relation sensitivity-confidence":

- Respect to SDIGA algorithm, the best interpretability is obtained with the KNNI approach for MVs, because it obtains the lowest number of rules and with a low number of variables per rule. With respect to the relation sensitivity-confidence is clearly presented in the results that the KNNI method obtains the best relation. Despite of the fact that KNNI does not obtain the best results in significance and unusualness, later is very similar to the best. All in all the best results are obtained with the KNNI method.
- In MESDIF is perfectly defined that the best method for the treatment of MVs as pre-processing step in this type of environments is the KNNI algorithm, because it obtains the best results in all the quality measures analysed.
- Finally, respect to NMEEF-SD algorithm it is more difficult to obtain a conclusive analysis. In interpretability the best results are obtained by the FKMI method. However, in the relation sensitivity-confidence and in interest the best results are obtained by the KNNI method. In summary, the best overall results are for KNNI, as this algorithm obtains the best results in the relation sensitivity-confidence, in significance, in sensitivity, in confidence and in interest, also obtaining a good interpretability, because the number of variables is low.

As can be observed in the analysis performed, the results indicate that in all the EFSs for SD task presented throughout the literature the best results are obtained using the KNNI method for the pre-processing of the data sets containing MVs. This statement is graphically depicted in Fig. 3, which presents the results obtained for each algorithm using the different MVs pre-processing methods. This figure presents a sub-figure for each algorithm, and verifies that the KNNI pre-processing method obtains the best overall results.

Some conclusions can be drawn according to the results of this experimentation:

- The KNNI pre-processing approach for MVs obtains the best results for subgroup discovery based on evolutionary fuzzy systems.
- The relation sensitivity-confidence is optimized when using the KNNI approach.
- The SD algorithm with the best relation sensitivity-confidence is NMEEF-SD, with values close to 100% in sensitivity and close to 80% in confidence.
- Respect to the interest guideline, SDIGA obtains the best results in significance while NMEEF-SD obtains the best results in unusualness.
- SDIGA obtains an excellent interpretability in this experimental study, with a low number of rules and variables.

5. Conclusions

In this paper a experimental study related to the effect of the use of missing data treatment in EFSs for SD is presented, where the most relevant imputation approaches for the treatment of MVs are used in order to preprocess some standard data sets.

It is very common that data sets have natural MVs. This problem leads to a loss of precision and quality of the results of the algorithms applied, because this type of values can not be covered. This problem is also present in the SD task, and it can be solved with the use of pre-processing MVs approaches.

The experimental study presented in this paper is performed with a number of data sets which contain both natural and induced MVs. Thus, the purpose was to apply SD algorithms in data sets preprocessed previously using MVs approaches in order to analyse which of them is the best option to preprocess data sets when using EFSs algorithms for the SD task. The results show that the
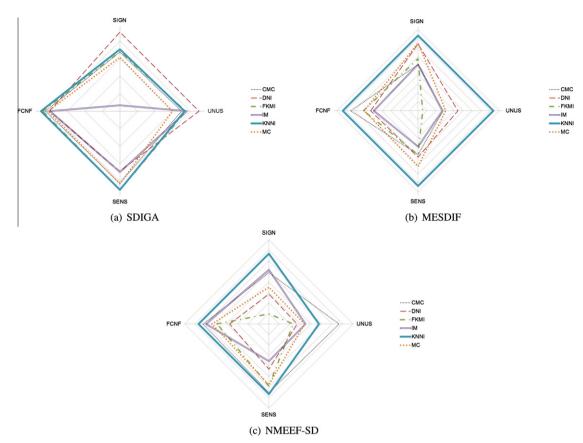
(a) SDIGA

(b) MESDIF

(c) NMEEF-SD

Fig. 3. Results obtained in the quality measures for Subgroup discovery: Significance (*SIGN*), unusualness (*UNUS*), sensitivity (*SENS*) and fuzzy confidence (*FCNF*).

most appropriated imputation method for MVs when used as a pre-processing step for the EFSs algorithm for SD is KNNI.

References

Acuna, E., & Rodríguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In *Proceedings of the meeting of the international federation of classification societies* (pp. 639–648).

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., et al. (2011). KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing, 17*, 255–287.

Alcalá-Fdez, J., Sánchez, L., García, S., Del Jesus, M., Ventura, S., Garrell, J., et al. (2009). KEEL: A software tool to assess evolutionary algorithms for data mining problems. *Soft Computing, 13*, 307–318.

Atzmueller, M., Puppe, F., & Buscher, H. P. (2004). Towards knowledge-intensive subgroup discovery. In *Proceedings of the Lernen–Wissensentdeckung–Adaptivität–Fachgruppe Maschinelles Lernen* (pp. 111–117).

Atzmueller, M., & Puppe, F. (2006). SD-Map – A fast algorithm for exhaustive subgroup discovery. In *Proceedings of the 17th European conference on machine learning and 10th European conference on principles and practice of knowledge discovery in databases* (pp. 6–17). Springer.

Barnard, J., & Meng, X. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research, 8*, 17–36.

Batista, G., & Monard, M. (2003a). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence, 17*, 519–533.

Batista, G., & Monard, M. (2003b). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence, 17*, 519–533.

Bay, S., & Pazzani, M. (2001). Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery, 5*, 213–246.

Berlanga, F. J., Del Jesus, M. J., González, P., Herrera, F., & Mesonero, M. (2006). Multiobjective evolutionary induction of subgroup discovery fuzzy rules: A case study in marketing. In *Proceedings of the 6th industrial conference on data mining* (pp. 337–349). Springer.

Carmona, C.J., González, P., Del Jesus, M.J., & Ventura, S. (2011). Subgroup discovery in an e-learning usage study based on Moodle. In *Proceedings of the international conference of European transnational education* (pp. 446–451).

Carmona, C. J., González, P., Del Jesus, M. J., & Herrera, F. (2009). An analysis of evolutionary algorithms with different types of fuzzy rules in subgroup discovery. In *Proceedings of the IEEE international conference on fuzzy systems* (pp. 1706–1711).

Carmona, C. J., González, P., Del Jesus, M. J., & Herrera, F. (2009). Non-dominated multi-objective evolutionary algorithm based on fuzzy rules extraction for subgroup discovery. In *Proceedings of the 4th international conference on hybrid artificial intelligence systems* (pp. 573–580). Springer.

Carmona, C. J., González, P., Del Jesus, M. J., & Herrera, F. (2010). NMEEF-SD: Non-dominated multi-objective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. *IEEE Transactions on Fuzzy Systems, 18*, 958–970.

Carmona, C. J., González, P., Del Jesus, M. J., Navío, M., & Jiménez, L. (2011). Evolutionary fuzzy rule extraction for subgroup discovery in a psychiatric emergency department. *Soft Computing, 15*, 2435–2448.

Deb, K., Pratap, A., Agrawal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation, 6*, 182–197.

Del Jesus, M. J., González, P., & Herrera, F. (2007a). *Fuzzy sets and their extensions: Representation, aggregation and models* (Vol. 220). Springer [Chapter. Subgroup Discovery with Linguistic Rules, pp. 411–430].

Del Jesus, M. J., González, P., & Herrera, F. (2007b). Multiobjective genetic algorithm for extracting subgroup discovery fuzzy rules. In *Proceedings of the IEEE symposium on computational intelligence in multicriteria decision making* (pp. 50–57). IEEE Press.

Del Jesus, M. J., González, P., Herrera, F., & Mesonero, M. (2007c). Evolutionary fuzzy rule induction process for subgroup discovery: A case study in marketing. *IEEE Transactions on Fuzzy Systems, 15*, 578–592.

Di-Nuovo, A. G. (2011). Missing data analysis with fuzzy C-means: A study of its application in a psychological scenario. *Expert Systems with Applications, 38*, 6793–6797.

Doh-Soon, K., & Kwang-Jae, K. (2012). A data mining approach considering missing values for the optimization of semiconductor-manufacturing processes. *Expert Systems with Applications, 39*, 2590–2596.

Dong, G. Z., & Li, J. Y. (2005). Mining border descriptions of emerging patterns from dataset pairs. *Knowledge and Information Systems, 8*, 178–202.

Eiben, A. E., & Smith, J. E. (2003). *Introduction to evolutionary computation*. Springer.

Farhangfar, A., Kurgan, L., & Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition, 41*, 3692–3705.

Farhangfar, A., Kurgan, L. A., & Pedrycz, W. (2007). A novel framework for imputation of missing values in databases. *IEEE Transactions on Systems, Man, and Cybernetics, Part A, 37*, 692–709.

García-Laencina, P., Sancho-Gómez, J., & Figueiras-Vidal, A. (2009). Pattern classification with missing data: A review. *Neural Computation & Applications, 9*, 1–12.

Gheyas, I. A., & Smith, L. S. (2010). A neural network-based framework for the reconstruction of incomplete data sets. *Neurocomputing, 73*, 3039–3065.

Gryzmala-Busse, J., & Hu, M. (2001). A comparison of several approaches to missing attribute values in data mining. *Rough sets and current trends in computing (RSCTC 2000), LNAI 2005* (pp. 375–385).

Grzymala-Busse, J., Goodwin, L., Grzymala-Busse, W., & Zheng, X. (2005). Handling missing attribute values in preterm birth data sets. In *Proceedings of 10th international conference of rough sets and fuzzy sets and data mining and granular computing* (pp. 342–351).

Hai, W., & Shouhong, W. (2009). Discovering patterns of missing data in survey databases: An application of rough sets. *Expert Systems with Applications, 36*, 6256–6260.

Herrera, F. (2008). Genetic fuzzy systems: Taxomony, current research trends and prospects. *Evolutionary Intelligence, 1*, 27–46.

Herrera, F., Carmona, C. J., González, P., & Del Jesus, M. J. (2011). An overview on subgroup discovery: Foundations and applications. *Knowledge and Information Systems, 29*, 495–525.

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. University of Michigan Press.

Hüllermeier, E. (2005). Fuzzy methods in machine learning and data mining: Status and prospects. *Fuzzy Sets and Systems, 156*, 387–406.

Kloesgen, W. (1996). Explora: A multipattern and multistrategy discovery assistant. *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence [pp. 249–271].

Kralj-Novak, P., Lavrac, N., & Webb, G. I. (2009). Supervised descriptive rule discovery: A unifying survey of constrast set, emerging pateern and subgroup mining. *Journal of Machine Learning Research, 10*, 377–403.

Lavrac, N., Flach, P. A., & Zupan, B. (1999). Rule evaluation measures: A unifying view. In *Proceedings of the 9th international workshop on inductive logic programming* (pp. 174–185). Springer.

Li, D., Deogun, J., Spaulding, W., & Shuart, B. (2004). Towards missing data imputation: A study of fuzzy *k*-means clustering method. In *Proceedings of 4th international conference of rough sets and current trends in computing* (pp. 573–579).

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data. Wiley series in probability and statistics* (1st ed.,). New York: Wiley.

Luengo, J., Sáez, J. A., & Herrera, F. (2012). Missing data imputation for fuzzy rule based classification systems. *Soft Computing, 12*, 863–881.

Matsubara, Edson T., Prati, Ronaldo C., Batista, Gustavo E., & Monard, Maria C. (2008). Missing value imputation using a semi-supervised rank aggregation approach. In *Proceedings of the 19th Brazilian symposium on artificial intelligence: advances in artificial intelligence* (pp. 217–226).

Pyle, D. (1999). *Data preparation for data mining*. Morgan Kaufmann.

Romero, C., González, P., Ventura, S., Del Jesus, M. J., & Herrera, F. (2009). Evolutionary algorithm for subgroup discovery in e-learning: A practical application using Moodle data. *Expert Systems with Applications, 36*, 1632–1644.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Siebes, A. (1995). Data surveying: Foundations of an inductive query language. In *Proceedings of the 1st international conference on knowledge discovery and data mining* (pp. 269–274). AAAI Press.

Wang, H., & Wang, S. (2010). Mining incomplete survey data through classification. *Knowledge and Information Systems, 24*, 221–233.

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *Proceedings of the 1st European symposium on principles of data mining and knowledge discovery* (pp. 78–87). Springer.

Wrobel, S. (2001). *Inductive logic programming for knowledge discovery in databases*. Springer [Chapter Relational Data Mining, pp. 74–101].

Yan, S., Yiyuan, T., Shuxue, D., Shipin, L., & Yifen, C. (2011). Diagnose the mild cognitive impairment by constructing Bayesian network with missing data. *Expert Systems with Applications, 38*, 442–449.

Yongsong, Q., Shichao, Z., Xiaofeng, Z., Jilian, Z., & Chengqi, Z. (2009). Estimating confidence intervals for structural differences between contrast groups with missing data. *Expert Systems with Applications, 36*, 6431–6438.

Yongsong, Q., Shichao, Z., Xiaofeng, Z., Jilian, Z., & Chengqi, Z. (2009). POP algorithm: Kernel-based imputation to treat missing values in knowledge discovery from databases. *Expert Systems with Applications, 36*, 2794–2804.

Zadeh, L. A. (1994). Soft computing and fuzzy logic. *IEEE Software, 11*, 48–56.

Zitzler, E., Laumanns, M., & Thiele, L. (2002). SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimization. In *International congress on evolutionary methods for design optimization and control with applications to industrial problems* (pp. 95–100).