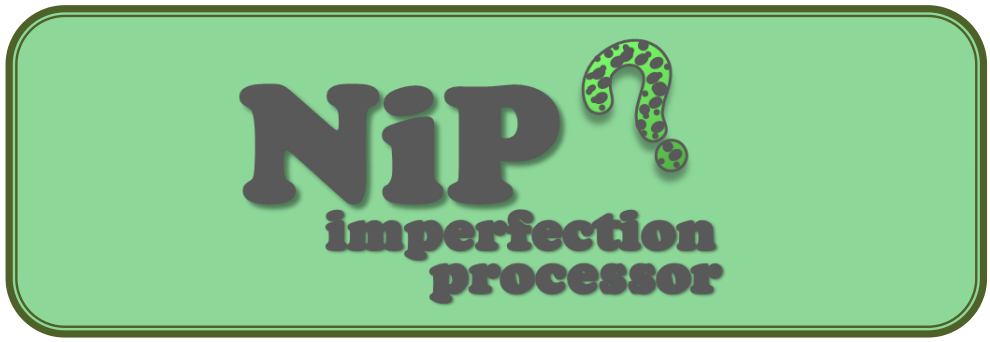
User's Guide to NIP Imperfection Processor



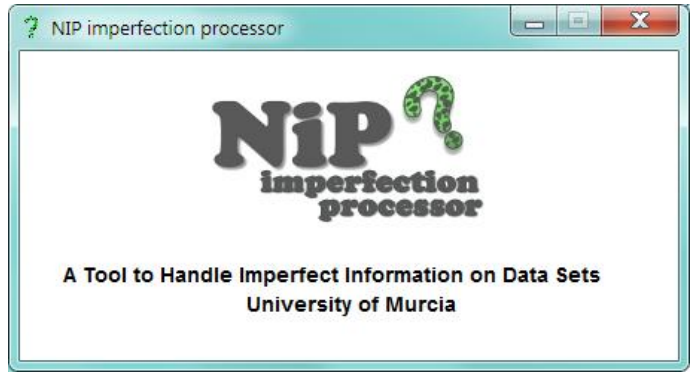
Faculty of Computer Science
University of Murcia. Spain 2012

José Manuel Cadenas Figueredo jcadenas@um.es
María del Carmen Garrido Carrera carmengarrido@um.es
Raquel Martínez España raquel.m.e@um.es
Emilio Serrano Fernández emilioserra@um.es

Contents

IMPORTANT NOTE:

NiP Imperfection Processor is a tool that allows us to generate and manage datasets with low quality data, created with the main purpose of being used in investigation due to absence of similar software that allows to establish a common framework for this kind of data. In this document, we show a fast guide to manage NIP tool.



It's important to note that this software is developed in Java 1.7. To execute this software it is necessary to have installed the version 1.7 or one later.

1. Set up the work directory and input dataset

Firstly, we need to select the work directory and to press the "Next" button. (Figure 1).

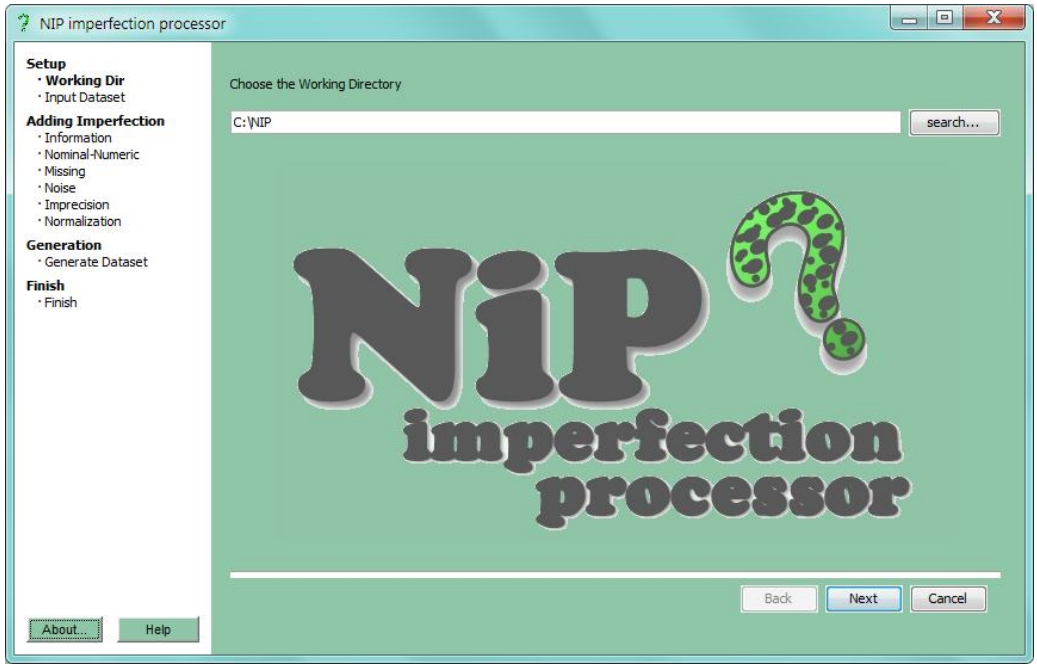


Figure 1

Later, we must indicate the dataset which we want apply changes, (Figure 2). After indicating dataset, we must select the format of it. If the format of dataset is a predetermined, we only select one of these formats otherwise we must define a custom format. If the input dataset has imprecise values, you must use a custom format, since the predefined formats do not allow this type of information.

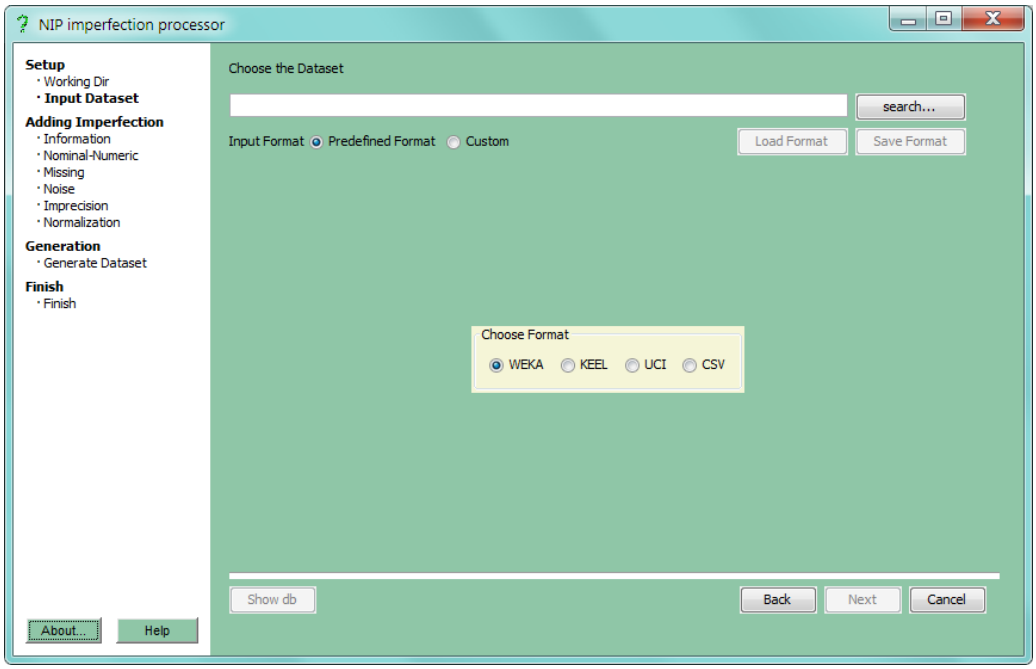


Figure 2

2. Adding Imperfection to dataset

Once we have selected the dataset and we have pressed the “Next” button, the information of the dataset is displayed on a table, (Figure 3). In this table we can see features of the attributes by means of following columns:

- Attribute column shows the name of the attribute. If the input format of dataset is Weka or Keel format, the names of the attributes in this column are the names indicate in the head of these formats. In other way, the names of the attributes are “atR*” to numerical attributes and “atN*” to nominal attributes (* is the corresponding number to the position of attribute in the dataset).

- Change Structure column indicates the position of attributes in the original dataset.
- Nominal-Numeric column shows if an attribute is the nominal or numerical.
- Missing column indicates if the attribute have missing values.
- Noise column shows if attributes have noise.
- Imprecision column indicates if attributes have fuzzy or interval values.
- Labels column indicates the number of possible labels that a numerical attribute can take if the imprecision of this attribute is introduced by means of fuzzy or interval partitions.
- Normalization column shows if an attribute has its values normalized.

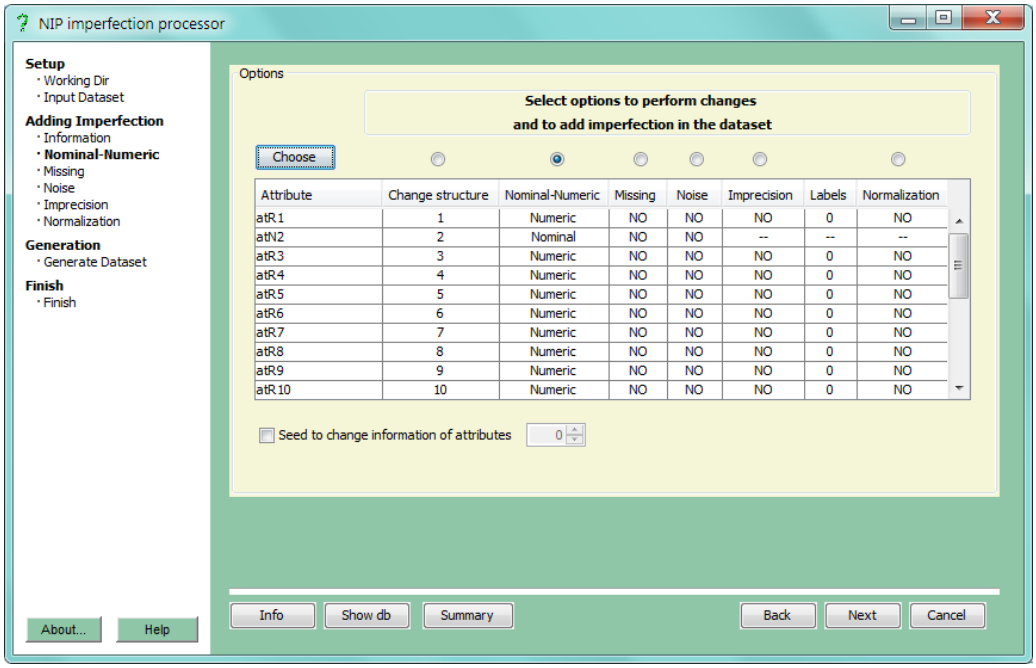


Figure 3

There is to take into account that the columns of imprecision, labels and normalization only show information about numerical attributes.

To modify the dataset we must select the option that we want (change structure, nominal-numeric, missing, noise, imprecision or normalization) and

to press the button choose. The modifications on the dataset will be displayed on the table in Figure 3.

In the different screens of the tool, the buttons Info, Show db and Summary are available. Info button displays some information about the attributes (as type, minimum, maximum, mean, percentages of missing, noise, fuzzy and interval values). Show db button displays the dataset and Summary button displays a summary table of the different types of imperfection of the dataset, indicating separately if this imperfection existed in the original dataset (columns IM-"Input Missing", IF-"Input Fuzzy", II-"Input Interval") or has been added (columns AM-"Added Missing", AN"Added Noise", FAP-"Fuzzy Added from a Partition", AFI-"Added Fixed Interval", ARI-"Added Random Interval", IAP-"Interval Added from a Partition"). This table is shown in Figure 4.

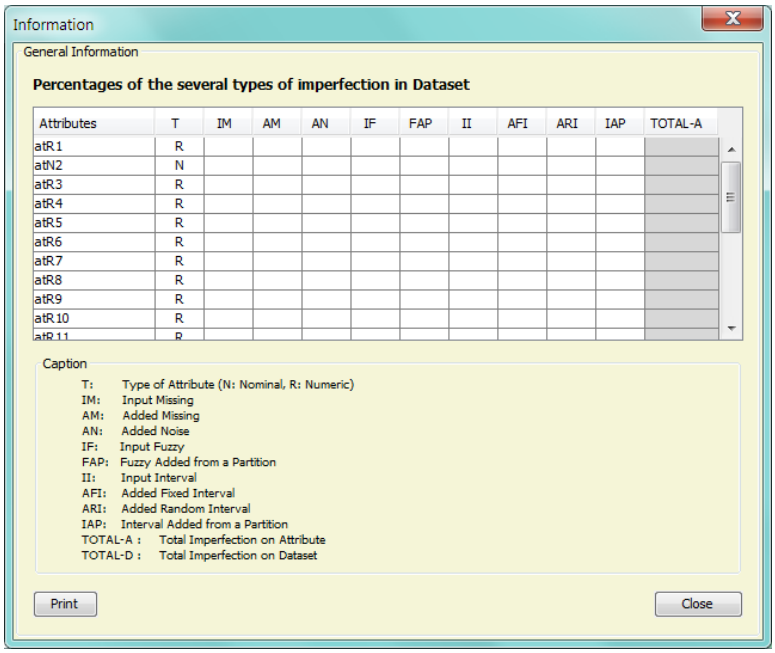


Figure 4

2.1 Change the structure of the dataset ("Change structure" option)

If we want to change the structure of the dataset, we select the option "Change structure" and we press the button "Choose". In this way we have the screen similar to Figure 5.

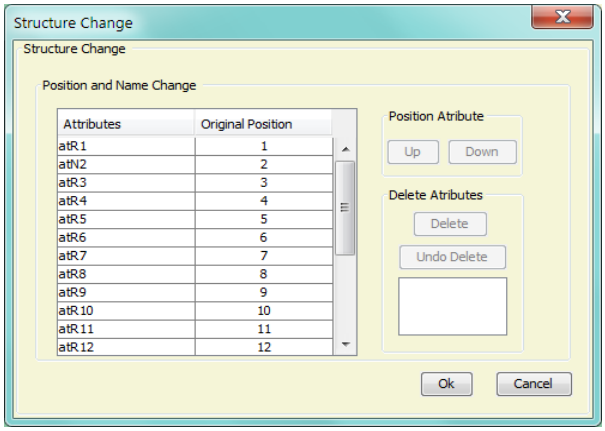
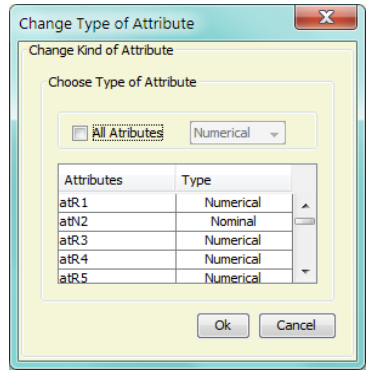


Figure 5

In the screen of change structure we can modify the position of the attributes, delete attributes or change the names of ones. To change the position of the attributes we select the attribute to change and we use the button “up” and “down” to modify its position. To delete attributes we select the attributes to delete and press the button “delete”. If we decide later not to delete an attribute, we select the attribute of the list of deleted attributes and press the button “undo delete”, and we recover this attribute again. To change the name of attributes we press the name box of the attribute and write the new name, but to confirm the new name we need to press the enter key.

2.2 Change the kind of attribute (“Nominal-Numeric” option)

If we want to change the kind of the attributes, we select the option “Nominal-Numeric” and we press the button “Choose”. In this way we have the screen similar to Figure 6.



In this screen, we can modify the kind of an attribute of two ways. On the one hand, if we want to modify all attributes to the same kind, we press the checkbox “All Attributes” and we select the kind of attribute

Figure 6

that we wish in the Combo Box. On the other hand, we can modify the kind of each attribute separately. For this, we press in the kind of the attribute that we want to change. It appears a Combo Box where we can select the kind that we want.

2.3 Adding missing values ("Missing" option)

If we want to add missing values to the attributes of the dataset, we select the option "Missing" and we press the button "Choose". In this case, it appears a screen similar to Figure 7.

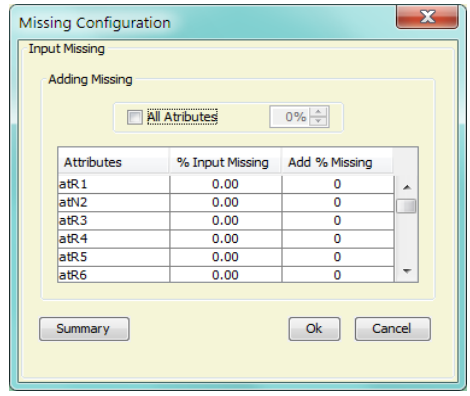


Figure 7

In this screen, we can add missing values to the attributes. The same way as change of kind of attribute we can treat this issue of two ways. One is to decide to add the same percentage of missing values of all attributes and the other is to decide to add missing values for each attribute. If we decide to modify all attributes at the same time, we select the checkbox "All Attributes" and we indicate in the spinner the percentage of missing for all attributes. Otherwise, we can write the percentage of missing in the column "Add % Missing" of each attribute. It's important to note that when we write the percentage of missing, we must press the enter key to confirm this percentage.

In both cases, the introduced percentage for each attribute will be added to percentage of missing that the attribute had in the original dataset (column "% Input Missing").

2.4 Adding noise ("Noise" option)

If we want to add noise to the attributes of the dataset, we select the option "Noise" and we press the button "Choose". In this case, we can see a screen similar to Figure 8.

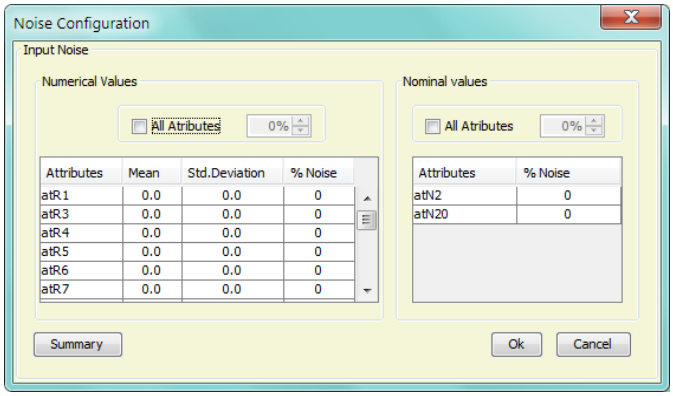


Figure 8

In this screen we can see two tables. The left table is to the numeric attributes and the right table to the nominal attributes. For each table, we have a different checkbox to add the same percentage of noise to all attributes. In this way we can add the same percentage to all numerical attributes and a different percentage to all nominal attributes or we can decide to add different percentage for each attribute. If we want this last option, we must write the percentage and press the enter key to confirm.

If we want to add noise to numerical attributes, this noise is Gaussian noise, that's why we must indicate a value of the mean and standard deviation for each attribute that we want to incorporate noise.

In the case of the nominal attributes, it is only necessary to express the percentage of noise to indicate the number of values of the attribute that will be changed randomly by another value of the domain.

2.5 Adding imprecision ("Imprecision" option)

If we want to add imprecision to dataset or we want to create a fuzzy or interval partition we must select the option imprecision and press the button "choose". The shown screen is similar to Figure 9.

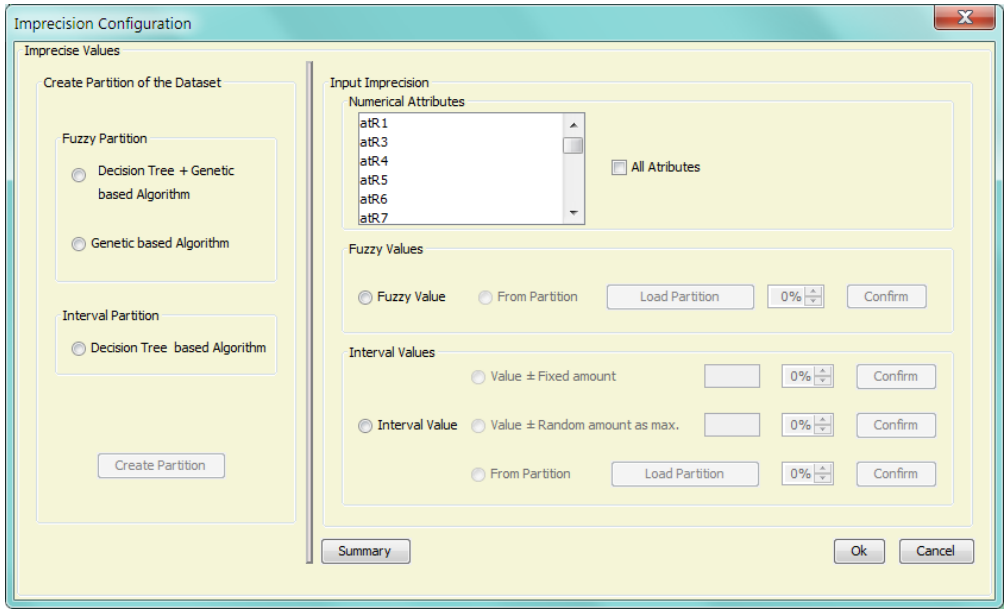


Figure 9

The screen of imprecision is divided in two parts. On the one hand, the left part is to create a fuzzy or interval partition without adding imprecision to the dataset. Algorithms that we can use to create a fuzzy partition are a decision tree and genetic based algorithm, [1], and a genetic based algorithm that is a version of the algorithm presented in [2]. If we want to create a crisp partition we can use a decision tree based algorithm which is used in the first phase of the algorithm presented in [1]. Note that each algorithm has a file of configuration to set the different parameters that they need. To set each parameter we can read the comments about it in the file. These files are located in the folder lib and their names are DT.config for the decision tree based algorithm, DT_GEN.config for the decision tree and genetic based algorithm and Gen.config for the genetic based algorithm. The partitions generated are stored in the working directory with the names "BD"DT_GEN_F.attrs, "BD"GEN.attrs and "BD"DT_C.attrs for the decision tree and genetic based algorithm, the genetic based algorithm and the decision tree based algorithm respectively.

On the other hand, in the right part of this screen we can add different kind of imprecision to the attributes. We can add fuzzy values from a partition which

we can create with the algorithms appointed before. In this screen, again we can add the same kind of imprecision to all attributes or for each attribute we set a different percentage of imprecision. If we add values of a fuzzy partition to all attributes, we must select the checkbox “All Attributes” and later we must select the percentage of imprecision in the active spinner and to press the button confirm.

When we press the button confirm about some option, the tool adds the kind of imprecision of the option with the percentage of imprecision that we have indicated to the attribute or attributes we have selected. Also, we can add interval values, which are created adding and subtracting a fixed amount to the value of the attribute “Value ± Fixed amount option” (always taking into account the domain of the attribute). Besides, we can add interval values, which are created adding and subtracting a random amount within the range [0,max] “Value ± Random amount as max. option” (again always taking into account the domain of the attribute). Also we can add interval values that we get when we create an interval partition with the algorithm based on decision tree. Options “Value ± Fixed amount” and “Value ± Random amount as max.” are mutually exclusive for the same attribute.

NOTE: We know that we have confirmed an option because when we press the button “confirm” the selected attributes are deselected.

2.6 Normalizing values (“Normalization” option)

If we want to normalize the values of the numerical attributes of the dataset, we select the option “Normalization” and we press the button “choose”. And it appears a screen similar to Figure 10.

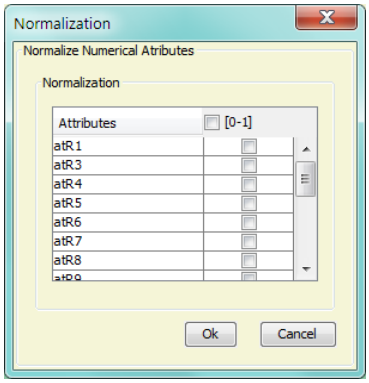


Figure 10

This option is only to numerical attributes and it use is similar to the other options, because in this case we can select to

normalize all attributes if we mark the checkbox of the head of the column “[0-1]”, or we can mark the check box for each attribute separately.

3. Making up the output format and Generating final dataset

Once, we have added imperfection in the dataset, we are going to set up the output format to get the transformed dataset. In this part of the tool, we can to decide the output format and if we want replace some values.

In a similar way for the input format we can get the dataset in a custom format or in a predefined format such as Weka, Keel, UCI or CSV. When we select a predefined format or a custom format and press de button “new”, we have a screen similar to Figure 11.

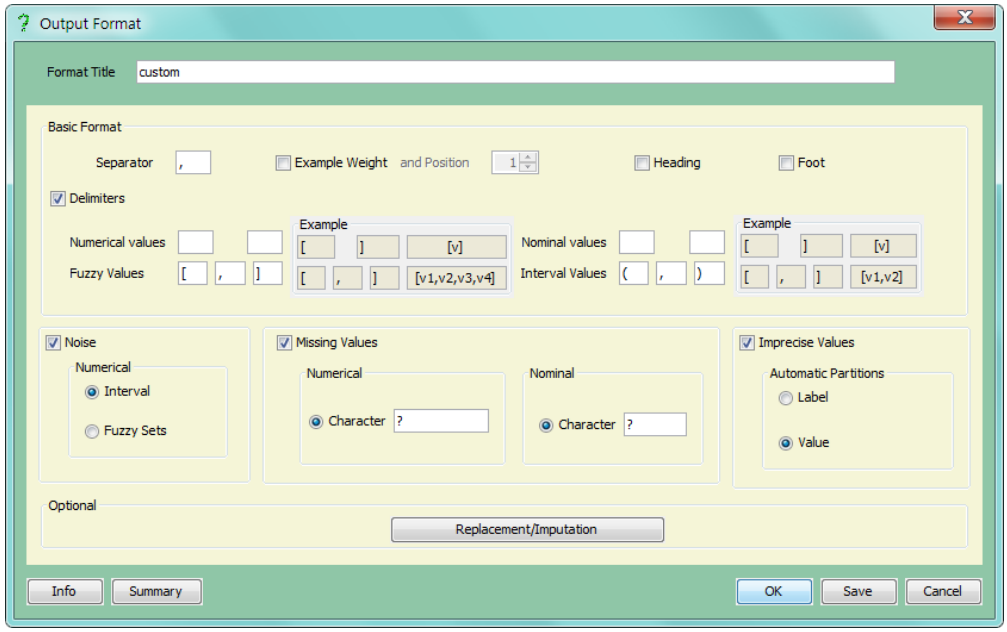


Figure 11

Figure 11 shows the default format of Weka format, because we have selected the option Weka. If we had select another option, we would have seen the default format of the option, and in the case of custom format, we would have seen all box with the default values because the user has to change these values.

Moreover to define the output format, in the screen of Figure 11, we can replace different values (nowadays we can only replace values but soon we may also impute values). If we press the button “Replacement/Imputation”, it’s appears a screen similar to Figure 12.

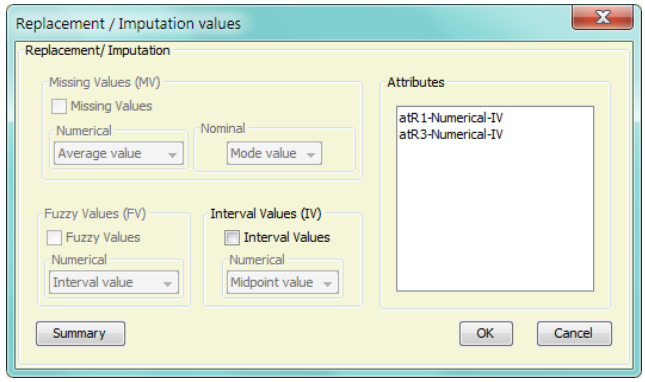


Figure 12

The screen of Replacement/Imputation will only show the activated options that the dataset has. For example, the dataset of Figure 12 only has interval values. Due to this, the interval option is the only activated. If we want to change interval values by other values, we must press the checkbox of interval values and select in the Combo Box the option that we want. Note that the name of the attribute is different because at the end of the name appears “-IV”. This suffix indicates that this attribute has interval values. If the suffix had been “-MV” or “-FV”, the attribute would have had missing or fuzzy values, respectively.

After defining the output format and replace values, we can do the same operation for the other formats because we can obtain the same dataset in different formats with different replacement at the same time.

The final datasets are in the work directory in a folder with the same name of the format. It’s important to take into account that the CSV, UCI and custom formats also create transformed dataset, they create a file with summarized information of the transformed dataset. This file has an extension “*.arff.spec”, “*.dat.spec”, “*.csv.spec”, “*.data.spec”, “*.custom.spec” to the transformed dataset in WEKA, Kell, CSV, UCI and custom formats, respectively.

4. Bug reports

Bug reports can be made to the following e-mail: raquel.m.e@um.es

References

[1] J.M. Cadenas, M.C. Garrido, R. Martínez-España, P.P. Bonissone, OFP CLASS: A hybrid method to generate Optimized Fuzzy Partitions for Classification, Soft Computing, 16 (4): 667-682, 2012

[2] Y-S. Choi, B.R. Moon, Feature Selection in Genetic Fuzzy Discretization for the Pattern Classification Problems, IEICE Transac, 90 (7): 1047–1054, 2007.