Machine Learning with Missing Attributes Values Methods Implementation

Štefánia GALLOVÁ, *Member, IAENG,* Michal AUGUSTIN, Sakena Saied Alsadig ALTAHR

Abstract—**One important and inconvenient problem is the presence of missing data in effort to achieve a data quality within our problem solving process. We notice the frequent occurrence of missing attributes values in real world data sets. There are some well- known strategies how to deal with missing value features within classification problem. At first, we apply five generally known investigated approaches to missing attribute values. Next, we use improved 6th approach to missing attribute values description and handling.**

Index Terms— **missing value, classification, control parameter, dataset, training set**

I. Introduction

This provides some experimental results for six missing attributes value approaches implementation. Classification with missing values is the problem of real world tasks. We have many cases when data sets attributes are not independent from each other. We have to identify the relationships among various attributes to achieve the missing values determination.

There are several strategies how to deal with missing features of value within classification problem solving. Various strategies combine the use of generative classification models, and the combination of standard discriminative methods with imputation, reduced models and response indicator augmentation.

Generally, we know the following approaches for description of missing attributes values.

1) *Most Common Attribute Value.* The attribute value that occurs most often is selected to be the value for all the unknown attribute values. This is a simple method. For example, the CN2 algorithm uses this approach.

2) *Concept Most Common Attribute Value Method (Maximum relative frequency method, or Maximum conditional probability method-given concept).* This is a restriction approach and does not pay any attention to the relationship between attributes and decision. The attribute value which occurs the most common within the concept is selected to be the value for all the unknown values of the attribute [3], [5]. *C4.5 method.* We consider entropy and the example splitting with missing attribute values to all concepts.

3) *All Possible Attribute Values Assigning Method.* We replace an example with missing attribute by a set of new examples, in which the missing attribute value is replaced by all possible values of the attribute. If example contains more that one unknown attribute value, at first – we will do our substitutions for one attribute, then – we will do the substitution for the next attribute, and so on, until all unknown attribute values are replaced by new known attribute values.[4].

4) *All Possible Values of the Attribute Restricted to the Given Concept Assigning Method.* This approach is not related to the concept. We have a restriction of the method of assigning all possible values of the attributes to the concept, indicated by an example with a missing attribute value.

5) *Ignoring Examples with Unknown Attributes Values Method..* This approach ignores the examples that have at least one unknown attribute value. The next step is using the rest of the table as input to the successive learning process realization.

6) *Event Covering Method.* This probabilistic method makes covering or selecting a subset of statistically interdependent events in the outcome space of variable-pairs, disregarding whether or not the variables are statistically interdependent.

7) *Missing Attribute Values as Special Values Treating Method.* We don´t try to find some known attribute value as its value. We use "unknown" itself as a new value for the attributes that contain missing values and treat it in the same way as other values [9].

8) A Method based on Special LEM2 Algorithm. This method omits the examples with unknown attribute values when building the block of attribute. We induce a set of rules by using the original LEM2 method.

II. PROBLEM SOLVING

The aim of our effort is to evaluate and compare the efficiency of some known effective algorithms for missing value handling implementation with our one suggested approach to missing value handling algorithm implementation. At first, we apply five generally known investigated approaches to missing attribute values. Next, we use better 6th approach to missing attribute values description and handling.

1) "M1" (*All Possible Attributes Values Assigning Method - restricted to the given concept)* This is an approach with concept indication by an example with a missing attribute value.

Štefánia GALLOVÁ, Associative Professor, Pavol Jozef Šafárik University in Košice, Faculty of Computer Science, + 421 948 158 082 e-mail: stefania.gallova@upjs.sk).

2) “M2” (*Most Common Attribute Value Concept Approach)* is the method which does not pay any attention to the judged relationship between attributes and the decision.
3) “M3” (*Ignoring Examples with Unknown Attributes Values Method.).* This approach ignores the examples that have at least one unknown attribute value. The next step is using the rest of the table as input to the successive learning process realization.
4) “M4” (Most Common Attribute Value Method) is the CN2 algorithm uses mentioned idea..
5) “M5” an *Imputation method with k-Nearest Neighbor* to estimate and substitute missing data. This method has some benefits, such as the following. It can predict qualitative attributes – “the most frequent value among the k nearest neighbors”, and quantitative attributes – “the mean among the k nearest neighbors” This is an imputation method. Concept named as imputation denotes a procedure that replaces the missing values in the data set by some plausible values.
6) “M6” is the 6th improved approach which provides better implementation results than above mentioned approaches. Some relevant aspects of this method are illustrated in the following chapter.

We solve the following novel classification approach. We use three factors for judging – specificity, strength and support. These factors decide about example to which concept it belongs. The definitions of introduced factors are the following. *Specificity* is the total number of rule attribute-value pairs on the rule left-hand side. The matching rules with a larger number of attribute-value pairs are considered more specific. *Strength* factor represents the total number of correctly classified examples .by the rule during process of training. *Prop* factor is defined as the sum of all matching rules scores from the judged concept. The concept *K* for which the prop, i.e. the expression (1) , is the largest is a winner, then the example is classified as being a *K* member.

$$\sum_{P} Strength(rule) * Specificity(rule) \qquad (1)$$

The meaning of concept *P is* “matching rules describing *K*”. If we have an example to be not completely matched by any rule , we use *partial matching* within classification system approach.

III. EXPERIMENTS

Table 1 illustrates input data files (i.e. used datasets named “Obrab.”,”Tvarn”,”.CNC”, respectively) in terms of the examples number, the number of concepts and the attributes number that describe the examples which were used in our experiments implementation. All three data files were taken from real-world environment where unknown attribute values more or less frequently occur. We realize also an artificial (synthetic) missing attributes values introduction with smaller level or higher level of value missing , that is expressed by graduating from 1 to 5th stage. The last stage represents the most amount of missing attributes introduction [2], [6], [7].

Table 1 illustrates the performance comparison for some real-world datasets, respectively. Generally, in each dataset, counts of positive class and negative class were almost the same. A baseline classifier would have the accuracy from 40% to 60% by classifying all the testing data points to be equal to 1 or 0. We reached the model with accuracy about 70%., i.e. the features that we had selected have appropriate discriminating ability. Datasets were obtained from real–world technological processes within knowledge base building process of solved domain expert system. Datasets contain relevant diagnostic knowledge for machine-tool equipments problems solving. We applied the most known classification techniques for prediction problem.

The average squared error mean is used as a metric tool to compare the performances. Our experience is that this metric is remarkably robust and has higher average correlation to other metrics. For this reason, it is more appropriate metric to compare of different classifiers than others. Average squared error finding in binary classification setup requires the posterior probability predicting instead of predicting just the class label. In fact, a model which could predict the true underlying probability for each test case would be optimal [1]. If we have an unbiased environment, the associated cost with the missing classification of positive and negative class is the same, and, in addition, no probability calibration is required. We have the following example. If the predicting probability value is above 0.5 and the judged sample is predicted as positive class. Then, the difference of 1 and the value is considered as the error. In contrast – if the value is below 0.5 – the sample is predicted as negative class and the difference of 0 and the value is considered to be the error. In the following, we suppose, that we have the worst case. We have an error value of 0.5. The label can be predicted only by tossing a fair coin. Then, a root mean squared error is computed over all the samples. This approach is used while computing the squared error.

At first, we examine if the proposed model is able to recover both the unknown selection mechanism and a correct model of the data. [2], [3]. We have the collaborative filtering domain for generating the synthetic data sets patterned after real data sets. We have parameter δ_v which models a value-based effect. Parameter γ_{ij} models a joint element index or latent effect of variable. This latter effect can include factors which are item-specific, i.e. a given data item I can have its own probability of being missing, and latent variable-specific, i.e. each mixture component j generates its own pattern of missing data. The values of these factors can be arbitrary real numbers and they are combined to obtain the selection probabilities through the logistic function as seen in equation 1. We can compute the complete set of the selection probability parameters μ_{vij} (according to equation (2), with given values for the model parameters δ_v and γ_{ij}.

$$\mu vij = (1 + e^{-(} \delta_v + {}^{\gamma}{}_{ij})^{-1} \qquad (2)$$

ISBN: 978-988-14047-2-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

We have multinomial mixture data model with 80 data variables, and the seven values per variable. We sample data cases from the mixture model to form a complete data set. We have the following equation $\mu_v(c) = c(v-5) + 0.5$. Introduced parameter c controls the strength of the missing data effect. We create 10 sets of observation probabilities by varying the parameter c from 0.00 to 0.10

The resulting parameters are illustrated in figures 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11. We have ten sets that contain the observation parameters. They were used to sample ten different training sets. We have ten different selections model that we construct with setting illustrated by the following way.

$$\delta_v (c) = \log (\mu v(c) / (1 - \mu v(c))) \qquad (3)$$

$$\gamma_{ij} = \log (u_{ij} / (1 - \gamma_{ij}))) \qquad (4)$$

Expression $\mu_v(c)$ is the result of the logistic function applying on δ_v (c). Expression u_{ij} is the result of the logistic function applying on γ_{ij} . We realize the computing of the corresponding set of selection probabilities $\mu_{vij}(c)$ and their using for sampling 10 different training sets. We have seven values per variable.

TABLE 1

DATASETS PERFORMANCE COMPARISON

Dataset	Training set size	Attributes number	Testing set size	Squared error
Obrab	420	19	55	0.211
Tvarn	350	17	59	0.196
CNC	480	16	43	0.151

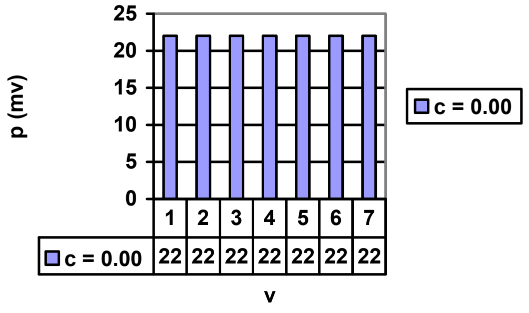


Fig.1. Selection probabilities (c=0.00).

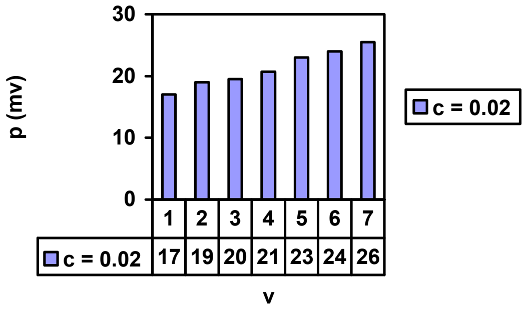


Fig.2. Selection probabilities (c=0.02)

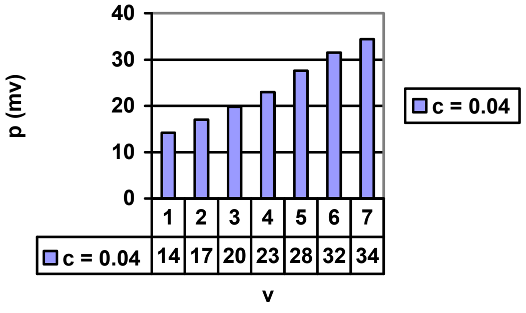


Fig 3. Selection probabilities (c=0.04)

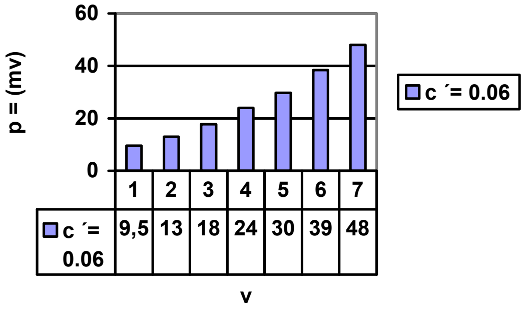


Fig.4. Selection probabilities (c=0.06)

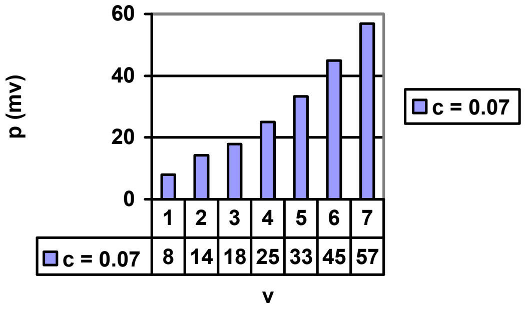
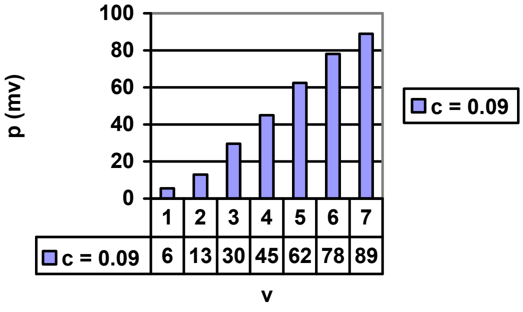


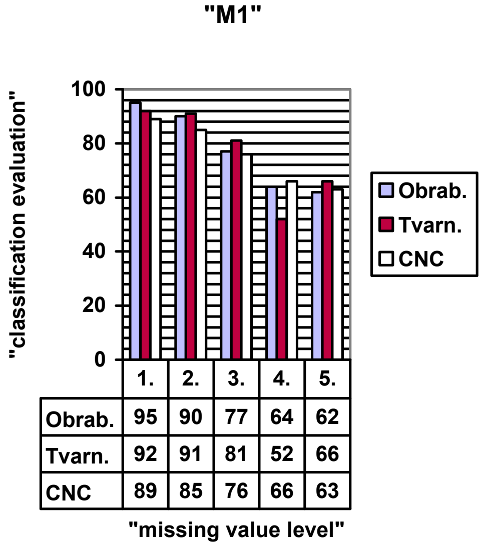
Fig 5. Selection probabilities (c=0.07)



	1	2	3	4	5	6	7
c = 0.09	6	13	30	45	62	78	89

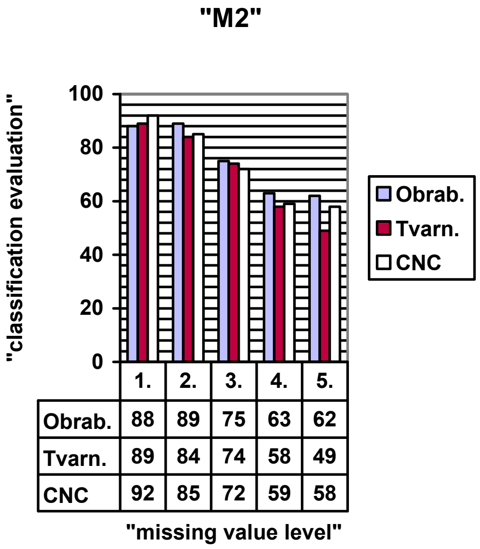
Fig 6. Selection probabilities (c=0.09)

Results of experiments with artificially implanted missing values indifferent rates and attributes into the data sets for six described methods are illustrated in the following pictures.



	1.	2.	3.	4.	5.
Obrab.	95	90	77	64	62
Tvarn.	92	91	81	52	66
CNC	89	85	76	66	63

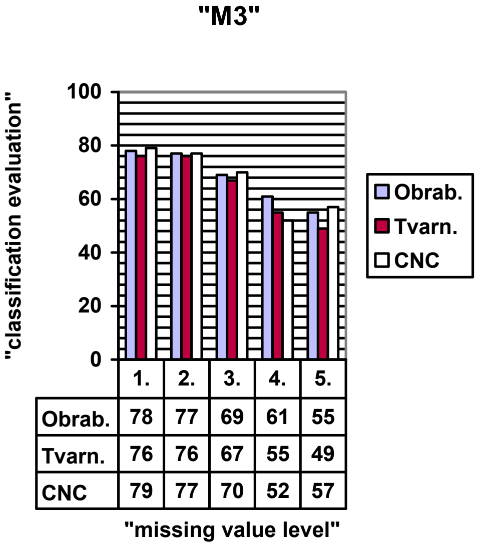
Fig.7. Implementation of " M1" method to missing attribute values handling. One column contains three united columns that represent results obtained for 3 datasets (Obrab., Tvarn., CNC) for certain missing attribute value level.



	1.	2.	3.	4.	5.
Obrab.	88	89	75	63	62
Tvarn.	89	84	74	58	49
CNC	92	85	72	59	58

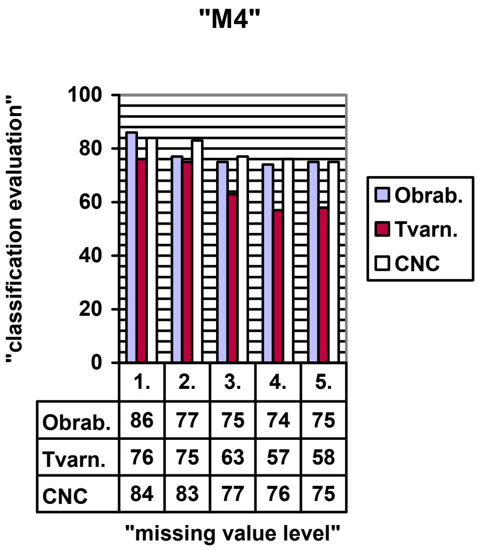
Fig.8. Implementation of " M2" method to missing attribute values handling. One column contains three united columns that represent results obtained for 3 datasets (Obrab., Tvarn., CNC) for certain missing attribute value level.

.



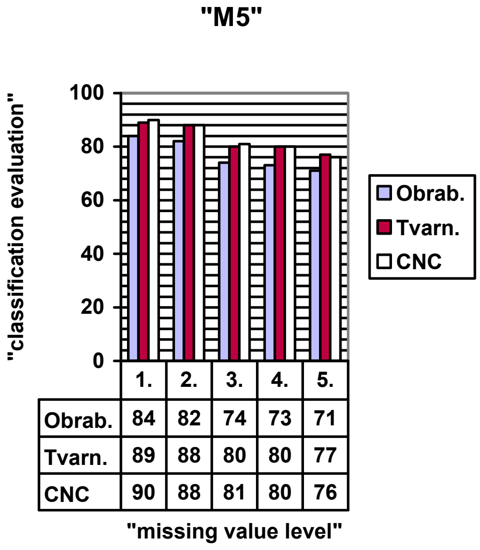
	1.	2.	3.	4.	5.
Obrab.	78	77	69	61	55
Tvarn.	76	76	67	55	49
CNC	79	77	70	52	57

Fig.9. Implementation of " M3" method to missing attribute values handling. One column contains three united columns that represent results obtained for 3 datasets (Obrab., Tvarn., CNC) for certain missing attribute value level.



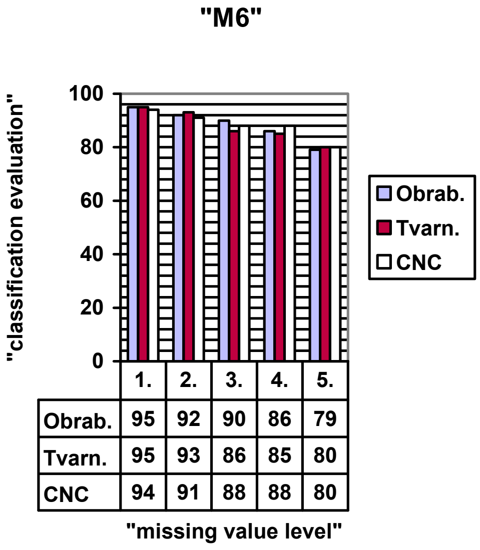
	1.	2.	3.	4.	5.
Obrab.	86	77	75	74	75
Tvarn.	76	75	63	57	58
CNC	84	83	77	76	75

Fig. 10. Implementation of " M4" method to missing attribute values handling. One column contains three united columns that represent results obtained for three datasets (Obrab., Tvarn., CNC) for certain missing attribute value level.



	1.	2.	3.	4.	5.
Obrab.	84	82	74	73	71
Tvarn.	89	88	80	80	77
CNC	90	88	81	80	76

Fig.11. Implementation of " M5" method to missing attribute values handling. One column contains three united columns that represent results obtained for 3 datasets (Obrab., Tvarn., CNC) for certain missing attribute value level.



	1.	2.	3.	4.	5.
Obrab.	95	92	90	86	79
Tvarn.	95	93	86	85	80
CNC	94	91	88	88	80

Fig.12. Implementation of " M6" method to missing attribute values handling. One column contains three united columns that represent results obtained for 3 datasets (Obrab., Tvarn., CNC) for certain missing attribute value level.

IV. CONCLUSION

Data quality is a major concern in algorithms of machine learning. There are other correlated areas, such as (first of all) data mining, knowledge and discovery from databases. We propose an approach how to deal with missing information presence effectively. This approach is better than comparable known models which do not account for the missing data mechanisms. Model implementation was applied on real-world collaborative sets of data. Proposed model do more better than comparable models that do not account for the missing data mechanism. We have introduced and verified learning and inference algorithms to jointly estimate the data and selection model parameters.

Used data sets results provide the fact that zero imputation and unconditional mean imputation rarely work well with linear classifiers implementation. With sufficient capacity, nonlinear classifiers can sometimes overcome poor imputation procedures. Multiple imputation was illustrated to work well so long as the conditional distribution of missing data given observed data is approximately correct. Some of the achieved results are the following. M5 method (k-nearest neighbour) has one following advantage. The missing attributes values handling is independent of the learning algorithm implementation. This advantage allows to have for each judged situation the selection with the most suitable imputation method. Our analysis and implementation of this approach provide the best results from the first four used known methods. Here is another advantage in connection with distribution of the complete data for any reason. The estimated error on observed test data can be an arbitrarily poor estimate of the error on the complete data. The second well-known method in our experiments was M4 (CN2). This approach is broadly used to missing attributes values handling.

The 6th method provides the best results within realized experiments. This approach is more sophisticated and robust than other ones.

REFERENCES

[1] B. Caruanna, A.Niculescu-Mizil, "Data mining on Metric Space:: An Empirical Analysis of Supervised Learning Performance Criteria" KDD, 319–352, 2004.

[2] P. Clark, T.Niblett,"The CN2 Inductiion Algorithm.." Machine learning vol 3., pp.261-283, 1989.

[3] S. Gallova, "Contribution to Icomplete and Noisy Information Problem Solving by Artificial Intelligence Principles Applying.", Lecture Notes in Engineering and Computer science, Nevswood Limited London, UK, 2010.

[4] J.W.Grzymala-Busse,"On the unknown attribute values in learning from examples.", Proc… of the ISMIS-91, 6th International Symposium on Methodologies for Intelligent systems, Charlotte, North Carolina, October 1991, Lecture Notes in Artificial intelligence, vol.542, Springer-Verlag, Berlin, Heidelberg, New York, pp.368-377, 1991.

[5] I.Knonenko, E.Roskar,"Experiments in automatic learning of medical diagnostic rules.", Technical Report, J.S.Institute, Ljubljana, 1984.

[6] B. Marlin, "Modeling user rating profiles for collaborative filtering.", Proceedings of the seventeenth Annual conference on Neural Information Processing systems (NIPS-2003), 2003.

[7] B.Marlin, "Collaborative filtering. A machiine learning perspective.", University of Toronto, January 2004.

[8] J.R.Quinlan,,"C4,5:Programs for Machine Learning.", Morgan Kaufman Publishers, San Matteo, CA, 1993.

[9] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.