medical data analysis,
missing data, data imputation,
classification efficiency

Tomasz ORCZYK[1], Piotr PORWIK[1]

INFLUENCE OF MISSING DATA IMPUTATION METHOD ON THE CLASSIFICATION ACCURACY OF THE MEDICAL DATA

Aim of this study is to show the dangers of filling missing data - particularly medical data. Because there are many dedicated medical expert systems and medical decision support systems, a special attention must be paid on the construction of classifiers. Medical data are almost never complete, and completion of the missing data requires a special care. The safest approach of dealing with missing data would be removing records with missing parameters and/or removing parameters that are missing in the records. Unfortunately reducing data set that is already very small is not always an option. Dangers coming out from data imputation are shown in the article, which presents the influence of selected missing data filling algorithms on the classification accuracy.

1. PROBLEM OVERVIEW

Common problem while analysing and classifying medical data is low number of medical records of patients with a disease of researcher's interest and also their incompleteness. This incompleteness may be due to various sources of data (due to the fact that different hospitals may have different sets of routinely performed tests and also this sets may vary over time). Usually some small subset of measured parameters is common to all patients, while other parameters may be randomly missing.

Problem of missing data is not new on the field of data analysis and statistics, there have been defined three categories of missing data [9],[11]: NMAR (not missing at random), MAR (missing at random) and MCAR (missing completely at random). Most of known research about data imputation methods assume that data is missing completely at random [10].

Also methods of dealing with the problem of missing data have been defined [10]:

- Instances discarding
  Remove rows containing missing value.
- Acquisition of missing values
  Re-acquire missing values; it is assumed that value is not missing, but has not been observed or measured and it is still possible acquire it.
- Imputation
  Replace missing values by a calculated value; inserted value can be based on a predictive model, distribution based model or can be a constant, unique value (i.e. when lack of observed value is also information).

[1] University of Silesia {tomasz.orczyk, piotr.porwik}@us.edu.pl.

- Feature reduction
  Remove the features that are missing from the data set or create separate classifiers for different missing features.

2. DATA CHARACTERISTICS

For the purpose of the experiment there have been used three datasets containing real medical data. This datasets have no missing data and have been used to train and cross-verify all tested classifiers. Based on this datasets there have been prepared 80 subsets of data with missing values introduced to them completely at random. For the first dataset there have been created 15 subsets with missing data distributed equally in rows (R) and 15 sets with missing data distributed equally in columns (C) and for remaining two sets there have been generated 10 sets with missing data distributed equally in rows and 15 sets with missing data distributed equally in columns. For the first data set, in each category, there were 5 sets with about 5%, 15% and 25% of missing fields (see Table 1 for exact amount of missing data fields), so in other words one category (R) represents patients which had marked 95%, 85% and 75% of all possible examinations (each patient's record is missing exactly the same number of parameters) and the other category (C) contains set where each parameter is marked for 95%, 85% and 75% of all patients (each parameter is missing foe exactly the same number of patient's records). For the next two datasets there was no subset with 5% of missing data fields for each patient as the had not enough parameters.

Table 1. Amount of missing values in tested data sets.

Subset	Effectively missing values [%]					
Database	C5	C15	C25	R5	R15	R25
HEPA	4.1	14.3	24.5	3.1	12.5	25
BREA	4.7	14.2	24.5	–	11.1	22.2
HEART	4.8	14.8	24.8	–	7.7	23.1
HEPA FULL	19.2					

All missing values were numeric parameters, the class label, which is an expert's diagnosis and was a decision variable for training classifiers was never missing, but it had some (unknown) level of uncertainty.

Data sets are as follows:

- HEPAtitis
  32 parameters, 49 records, 4 classes
  Classes distribution: 8%, 16%, 27%, 49%.
- BREAst tissue [7]
  9 parameters, 106 records, 6 classes
  Classes distribution: 13%, 14%, 15%, 17%, 20%, 21%.
- HEART disease [4]
  13 parameters, 270 records, 2 classes
  Classes distribution: 44%, 56%.
- HEPAtitis (full) [8]
  37 parameters, 118 records, 4 classes
  Classes distribution: 8%, 21%, 33%, 37%.

3. RESEARCH ENVIRONMENT

Main part of the experiment has been performed using a KNIME [1] environment with classifiers from the WEKA [6] tool. KNIME is an innovative visual programming environment for implementing

and testing algorithms, focused mainly on data mining.
For the purpose of introduction of missing data to the datasets a dedicated C# application has been written. This simple console .NET application operate on CSV formatted data and randomly removes given amount of values in each column or row.

Missing data filling methods:

- Globally average value
  This is the most trivial method, proposed by default by various data analysis tools like KNIME or Matlab. The average value of a given parameter is calculated from all records with this parameter defined.
- Average value within a cluster
  Data is clustered into n-clusters, where n is the number of decision variable classes. In this example an Expectation Maximization clustering (unsupervised learning) algorithm has been used.
- Average value within a class
  Parameter average value is calculated within a decision class.

All tested data imputation methods were based on a predictive model (namely – on arithmetic mean value).
Class average method cannot be used in final classification system, as it cannot fill missing values in unclassified data, so it is only useful during the testing stage of building a classification system, however two remaining methods can become a part of a classification system in its final shape.
All classifiers that have been used can work with missing data and obtained results proves that they achieve results closest to the original data without using any of tested missing data imputation methods, thus it is possible to compare results obtained from the data set that is naturally missing some values.
For the purpose of the experiment a workflow has been created using the KNIME environment with WEKA data analysis package. The workflow contained a main loop iterating through data set files and four branches of classifiers: without filling missing data, with globally average value filling method (GA), with cluster average value filling method (EM) and with class average value filling method (CA).
Classification accuracy has been verified using a "leave one out" cross validation method. All tested data imputation methods were single data imputation methods, what means that records has not been multiplied by the algorithm.
Following classification methods have been tested in the experiment:

- Naive Bayes [5]
- Random Trees [3]
- Random Forests [2]

As a measure of quality of missing data imputation algorithm a difference between overall classification accuracy on original and filled data has been used – closer to 0 is better. Either gain or decrease of classification accuracy on imputed data is unwanted.

4. THE RESULTS

Figures 1–2 illustrate obtained results. Line marked *0%* represents classification accuracy for original dataset (without missing values), bars represents the average classification accuracy for datasets with missing data. For every *n%* R and *n%* C dataset class there were 5 different datasets with the same amount of data missing and each bar in the graph represents an average classification accuracy for these 5 subsets of data.

The last figure (Fig. 4) represents results of the test made on the original (full) hepatitis dataset. There is no *0%* reference accuracy as this data set is naturally missing some data, but basing on the previous results it can be assumed that classification results on data without imputation can be used as a reference point. Obtained results are similar to results on reduced version of this database, but

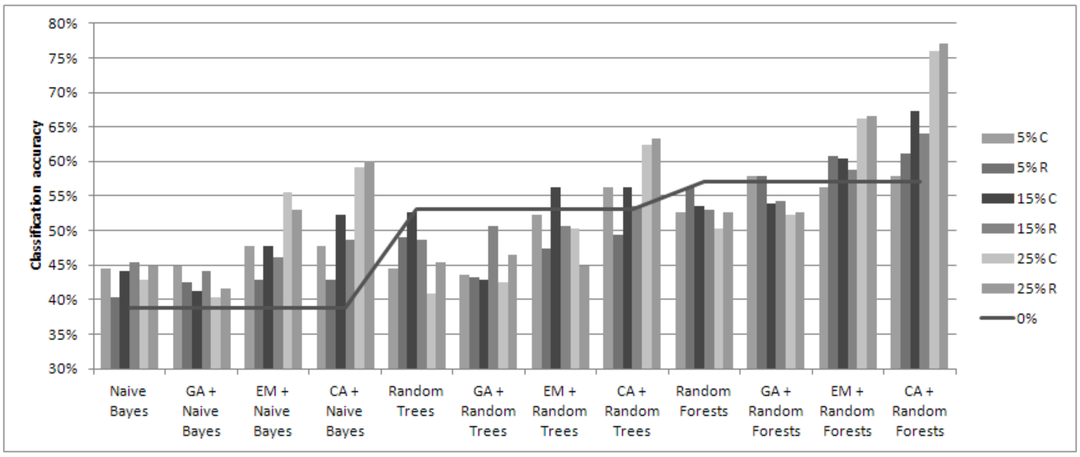

Fig. 1. HEPAtitis database results.

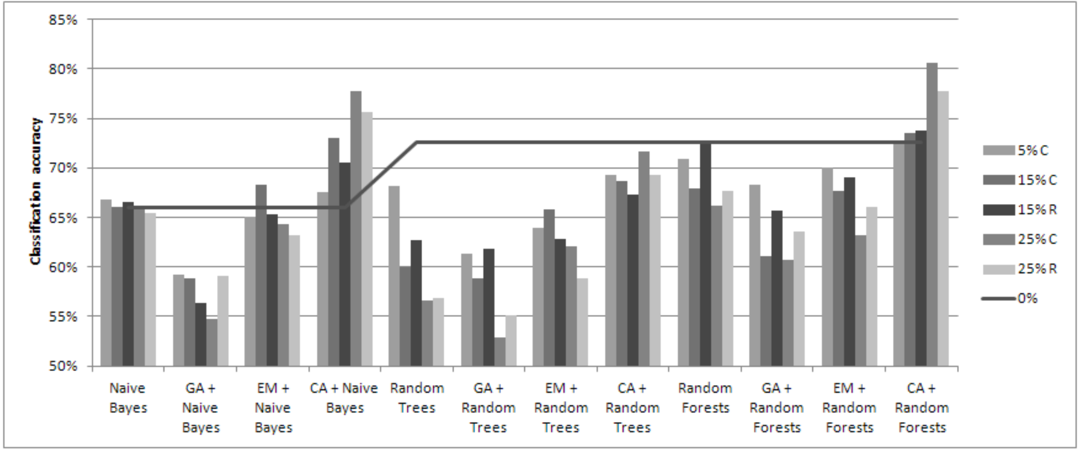

Fig. 2. BREAst tissue database results.

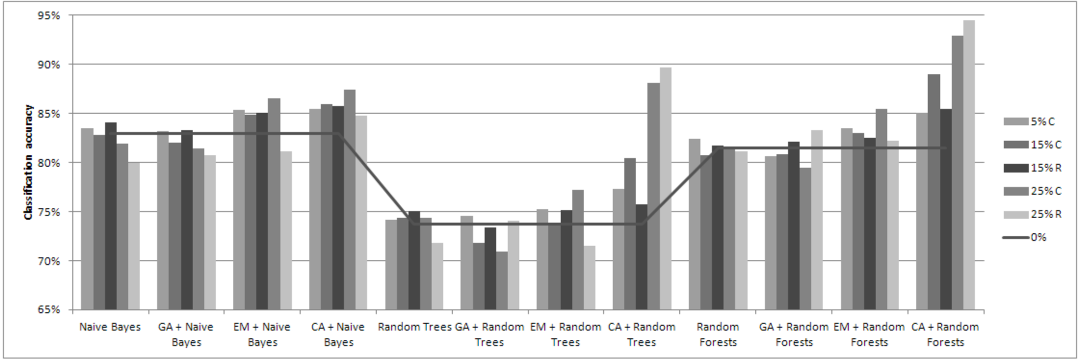

Fig. 3. HEART disease database results.

accuracy of classification without imputation and using global average values and using cluster average values are much closer to each other. Class average value imputation fails, same as in the case of 25% R/C Hepatitis subsets. Table 2 illustrates average classification accuracy of all 3 tested classifiers for each database and for each missing data imputation method with additionally calculated classification accuracies for subsets containing low amunt of missing data (∼5%) and high amount of missing data (∼25%). From the graphs above and Table 2 it can be seen that there is no universal method for datasets with small amount of missing data and for sets with high amount of missing data. For 5%R/C subsets of HEPA dataset the best missing data imputation method was global average value imputation and the worst was imputation of class average value imputation, while for 5%C subsets of BREA global average value imputation was the worst method and the best method was imputation of class average value.

However, if imputation of missing data is about to be a step in the data preprocessing of a final

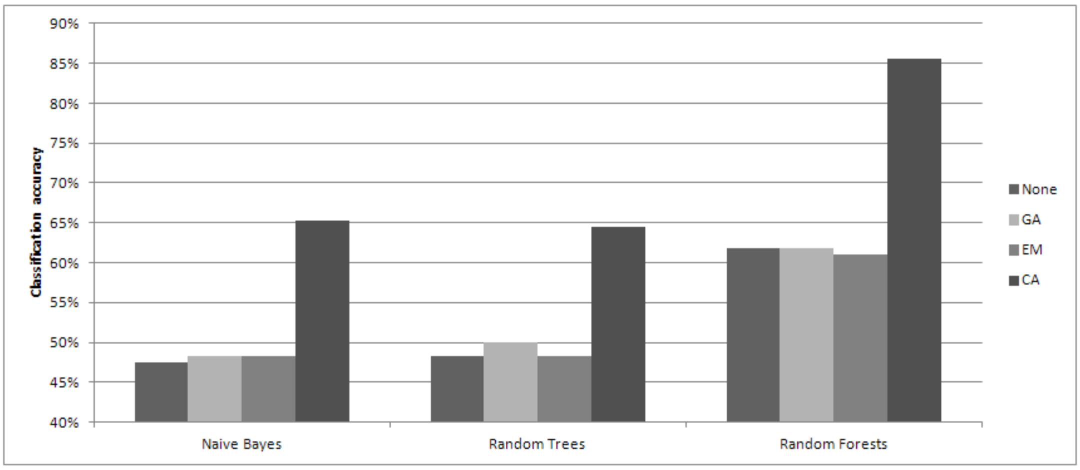

Fig. 4. Full HEPAtitis database results.

classification system and the percentage of missing data is low (∼5%), it is worth to try to cluster the data and impute average value within each cluster rather than impute globally average value.

It is also worth to notice that in most cases class average method has a tendency to overestimate the results while the global average method has a tendency to underestimate the results.

Table 2. Summary of classification accuracy.

Database	Fill method	Aver [pp]	5% [pp]	25% [pp]
HEPA	None	-2	-2	-4
	Global Av.	-2	-1	-4
	Clust. Av.	4	2	6
	Class Av.	9	3	17
BREA	None	-5	-2	-7
	Global Av.	-11	-7	-13
	Clust. Av.	-5	-4	-7
	Class Av.	2	-1	5
HEART	None	0	1	-1
	Global Av.	-1	0	-1
	Clust. Av.	1	2	1
	Class Av.	6	3	10
Av. Abs.	None	2	1	4
	Global Av.	4	3	6
	Clust. Av.	4	3	5
	Class Av.	6	2	11

5. CONCLUSIONS

Besides presented relationships between missing data imputation method and classification accuracy one important thing can be noticed: the scale of under– or overestimation of classification accuracy that can be caused by choosing wrong method. The miss-estimation of classification accuracy can exceed 10 percentage points.

There was no significant difference between two tested patterns of missing data (equal distribution of missing data fields in rows or in columns). Also results obtained with full Hepatitis data set (where data is missing at random but not completely at random) resemble results obtained from it’s subset containing only complete records, to which missing data was introduced completely at random.

Research has been done on a limited number of datasets and limited number of missing data imputation methods has been tested, so further research may be required to confirm obtained results.

BIBLIOGRAPHY

[1] BERTHOLD M. R., CEBRON N., DILL F., GABRIEL T. R., KÖTTER T., MEINL T., OHL P., SIEB C., THIEL K., WISWEDEL B., KNIME: The Konstanz Information Miner, Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Springer, 2007.

[2] BREIMAN L., Random Forests, Machine Learning, 2001, Vol. 45(1), pp. 5–32.

[3] DIETTERICH T. G., An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization, Machine Learning, 2000, Vol. 40(2), pp. 139–157.

[4] FENG C., SUTHERLAND A., KING R., MUGGLETON S., HENERY F., Comparison of machine learning classifiers to statistics and neural networks, Proceedings of the Third International Workshop in Artificial Intelligence and Statistics, 1993, pp. 41–52.

[5] FRIEDMAN N., GEIGER D., GOLDSZMIDT M., PROVAN G., LANGLEY P., SMYTH P., Bayesian Network Classifiers, Machine Learning, 1997, pp. 131–163.

[6] HALL M., FRANK E., HOLMES G., PFAHRINGER B., REUTEMANN P., WITTEN I. H., The WEKA Data Mining Software: An Update, SIGKDD Explorations, 2009, Vol. 11(1).

[7] JOSSINET J., Variability of impedivity in normal and pathological breast tissue., Med. & Biol. Eng. & Comput, 1996, Vol. 34, pp. 346–350.

[8] KRAWCZYK B., WOŹNIAK M., ORCZYK T., PORWIK P., MUSIALIK J., BŁOŃSKA-FAJFROWSKA B., Classification techniques for non-invasive recognition of liver fibrosis stage, Journal of MIT, Vol. 20, 2012, pp. 121–127.

[9] LITTLE R. J. A, RUBIN D. B., Statistical Analysis with Missing Data, New York: John Wiley & Sons, 1987.

[10] SAAR-TSECHANSKY M., PROVOST F., CARUANA R., Handling missing values when applying classification models, Journal of Machine Learning Research, 2007, Vol. 8, pp. 1217–1250.

[11] SCHAFER J. L., Analysis of Incomplete Multivariate Data, Chapman and Hall/CRC, 1997.