Artif Intell Rev (2013) 40:231–245
DOI 10.1007/s10462-011-9282-2

Missing values: how many can they be to preserve classification reliability?

Martti Juhola · Jorma Laurikkala

Published online: 21 August 2011

Abstract Using five medical datasets we detected the influence of missing values on true positive rates and classification accuracy. We randomly marked more and more values as missing and tested their effects on classification accuracy. The classifications were performed with nearest neighbour searching when none, 10, 20, 30% or more values were missing. We also used discriminant analysis and naïve Bayesian method for the classification. We discovered that for a two-class dataset, despite as high as 20–30% missing values, almost as good results as with no missing value could still be produced. If there are more than two classes, over 10–20% missing values are probably too many, at least for small classes with relatively few cases. The more classes and the more classes of different sizes, a classification task is the more sensitive to missing values. On the other hand, when values are missing on the basis of actual distributions affected by some selection or non-random cause and not fully random, classification can tolerate even high numbers of missing values for some datasets.

Keywords Medical data · Missing values · Distance measures · Imputation · Classification · Nearest neighbour searching

1 Introduction

For various reasons some variable values can be missing in a dataset. Missing values may frequently occur in numerous application areas. In particular, medical datasets often contains missing values. Therefore, we used medical datasets in our tests. Perhaps a physician who investigates patients deems some tests or questions unimportant for patients of a certain type. For instance, he or she has a good hypothesis about a diagnosis and, therefore, ignores variables like some medical tests that would not be important for that diagnosis. Some variables may concern diagnostic tests such as some medical imaging types, e.g. positron emission tomography (Mykkänen et al. 2000), that can be complicated or expensive to perform and,

M. Juhola (✉) · J. Laurikkala
Department of Computer Sciences, University of Tampere, 33014 Tampere, Finland
e-mail: Martti.Juhola@cs.uta.fi

thus, rarely accomplished. Perhaps there is not precise enough data, e.g., about the beginning of a symptom. Simply, some variables may remain unknown because their values could not be acquired or were noticed to be incorrect.

Several works considering missing values (Little and Rubin 1987; Pyle 1999; Witten and Frank 2000) exist, especially how to handle such problems, but research on the influence of the quantities of missing values are mostly missing and particularly regarding medical datasets. Usually missing values are substituted or imputed with artificial values that are suitable estimates computed from the known variable by variable in each class if there are appropriate training data available. This supervised situation requires that the class labels of all training data are known. If only "unknown" or new data without class labels is available, this cannot be performed classwise, but in an unsupervised manner from the whole dataset at a time. Naturally, estimates of the latter situation are less representative. If there are missing values in a training set, in principle such training cases could be abandoned entirely. If a considerable number of the variable values of an individual case were missing, this might be a reasonable choice. If there were only few missing values in a case, this would be unreasonable. Besides, there are such datasets, medical in particular, that relatively few values are missing, but most of the cases include at least one missing value. Thus, this straightforward elimination of cases is frequently impractical or even impossible. Instead, we need efficient techniques to substitute or to impute missing values.

An occasional reason for missing values in medical datasets is that there is no reason to store a normal value as a value of healthy people. In such cases these "normal" or "assumed" values would be a possibly good for imputation. However, this reason is not always known for a variable or there are other reasons for missing values (Chowdhury et al. 1991). Simple and fast methods to compute values to be imputed are to use central values as medians or means for quantitative variables and modes for binary and nominal (qualititative) variables. More sophisticated techniques include the use of linear regression or expectation-maximization (EM) algorithm (Pyle 1999), or nearest neighbour searching or neural networks (Pesonen et al. 1998). On the other hand, although the EM algorithm is often seen to be effective, the problem of such techniques is the requirement of statistical model assumptions (Fortes et al. 2006), e.g. about data distributions, which are not easy to satisfy or even to discover when a dataset is not known in advance. Using nearest neighbour searching (Wasito and Mirkin 2005) the closest or most similar case with a known variable value is computed and this known value is used to replace the missing value of the same variable. Sometimes even random substitution may be relevant, for instance if there are rather few missing values or they are missing in less important variables. On the basis of the earlier studies (Pesonen et al. 1998; Laurikkala et al. 2001), we can assume that usually the imputation technique applied affects only slightly classification accuracy. For instance, the use of means or neural network outputs yielded average accuracies of 78%, random values and nearest neighbour searching gave average accuracies of 74% and "normal" values those of 70 % for acute abdominal pain data (Pesonen et al. 1998).

Impact of missing values in datasets has seldom been investigated for the purpose to evaluate their influence on classification results. Using neural networks such influence was recently researched (Markey et al. 2006) for mammographic breast image and other associated data. Nevertheless, they applied complete training data, i.e. no missing values in training sets. After all, this is a restricted situation, which is often not met in reality.

In the present article we report our results after researching the effects of increasing numbers of missing values on the true positive rates and accuracies of classification tasks in five medical datasets. Two straightforward ways were employed to handle missing values and three basic methods were utilized for classification.

2 Datasets

We utilized three datasets from the UCI data repository (Blake and Merz 1998) frequently used in machine learning studies and two from our own previous research (Laurikkala and Juhola 1998; Viikki et al. 1999). The three former are Buba, Haberman and New-thyroid, and the two latter Incontinence and Vertigo.

The Buba dataset contains data about liver disease male cases. It consists of six variables and two classes as ‘sick’ or ‘healthy’ with 145 and 200 cases. The Haberman dataset includes breast cancer data of three variables. Its two classes express whether a patient died during a five-years period after surgery. There were 225 ‘survivors’ and 81 ‘dead’. The New-thyroid dataset consists of data concerning about the functioning of the thyroid. It has three classes: ‘normal function’, ‘hyperfunction’ and ‘hypofunction’ with 150, 35 and 30 cases. All variables of the three datasets are quantitative (ordinal or real). The datasets have 345, 306 and 215 cases.

The Incontinence dataset incorporates the data of female urinary incontinence cases with 13 binary variables. There are five classes: ‘stress’, ‘mixed’, ‘sensory urge’, ‘motor urge’ and ‘normal’ with 323, 140, 33, 15 and 18 cases. The total number of cases is 529. The vertigo dataset contains 20 ordinal, 16 real variables and 14 binary variables. It has six classes: ‘vestibular schwannoma’, ‘benign positional vertigo’, ‘Menière’s disease’, ‘sudden deafness’, traumatic vertigo’ and ‘vestibular neuritis’ with 130, 146, 313, 41, 65 and 120 cases, respectively. There are 815 cases altogether.

3 Methods

We applied 1-nearest neighbour searching by implementing a one-leave-out test procedure in Matlab for each dataset. In other words, a test set only included one test case at time and the other n-1 cases were used for its training set. Thus, n tests were executed for each test condition. This is a reasonable choice when datasets are relatively small and especially when they include small classes of few cases. We randomly marked values of variables as missing, but set the total number of the missing to be 10, 20, 30% etc. from the whole. We ran these tests with two ways to handle missing values: a suitable distance measure called the heterogenous Euclidean-overlap metric (Aha et al. 1991; Wilson and Martinez 1997) and imputation with the medians and modes of variable values. After choosing the latter of them, we continued by applying k-nearest neighbour searching with k equal to 3 and 5, linear discriminant analysis and naïve Bayesian classification to compare their classification results. Since the preceding tests were performed after fully randomly marking values as missing, we proceeded with 3-nearest neighbour searching, when values were first marked randomly as missing and according to classwise variable distributions of missing values present in the Incontinence and Vertigo datasets.

At first, we made two main test series with all five datasets to find out whether they yield different results while considering missing values differently. We handled missing values applying the heterogeneous Euclidean-overlap metric (HEOM). This distance measure is appropriate to medical data which often consists of mixed-type variables, i.e. from the simplest binary type to quantitative variables. It also takes into account missing values while calculating a distance. If one or both values of a variable of two cases are missing, the distance between the cases is defined to be the normalized maximum 1. Alternatively, we imputed missing values employing medians of quantitative variables and modes of binary variables (Little and Rubin 1987). This approach can be seen as the most straightforward imputation technique in addition to the use of random values. Modes and medians were computed variable by variable from the existent values before the imputation. The Euclidean metric was then applied to nearest neighbour searching.

HEOM is as follows (Aha et al. 1991; Wilson and Martinez 1997). Let us have patient cases x and y when x_i, and y_i are the values of a variable $i, i = 1\ldots, m$. A distance $D(x, y)$ between the cases is as follows.

$$D(x, y) = \sqrt{\sum_{i=1}^{m} d(x_i, y_i)^2}$$

For a qualitative or nominal (only binary in our datasets) variable a distance value $d(x_i, y_i)$ is

$$d(x_i, y_i) = \begin{cases} 1, & x_i \neq y_i \\ 1, & x_i \quad \text{or} \quad y_i \text{ missing} \\ 0, & x_i = y_i \end{cases}$$

and for a quantitative variable (ordinal, interval or ratio scales) a distance value is assigned with

$$d(x_i, y_i) = \begin{cases} \frac{|x_i - y_i|}{R_i} \\ 1, x_i \quad \text{or} \quad y_i \text{ missing} \end{cases}$$

in which R_i is the range of the variable i used to normalize the absolute difference.

The maximum distance was assumed if one or both variable values were missing. This is the so-called "pessimistic" way assuming that the difference is then as great as possible concerning the variable of a missing value. The "optimistic" way would define to be the minimum distance. The use of variable medians and modes is their "in-between" alternative relying on central values as estimates. This sounds a good approach, but its weakness is that they can be precisely estimated only along with classwise values, i.e. for a training set, where class labels are known. For unsupervised classification and always for a test set this is not possible, but they have to be estimated by applying the whole training (or test) set at a time. This is sometimes a clear weakness, because their actual values can vary from a class to class considerably. In fact, this variation is the key point of any classification.

Second, as mentioned earlier, we tested with k-nearest neighbour searching employing only the imputed versions of the data sets. Thirdly, we executed linear discriminant analysis and naïve Bayesian classification. Fourthly, we still studied the influence of actual missing value distributions when missing values were increased on the basis of these distributions in the Incontinence and Vertigo datasets. Some variables can have more missing values than others. Some may frequently contain missing values, but others do no include them at all. These circumstances may also differ between classes.

4 Tests and results

For the first test condition of no missing value where there is no randomness the tests were run once. For all other, 100 test runs were repeated to average any possible variation due to the random selection of missing values. Test conditions were 0, 10, 20, …, 90% of missing values from the total. Of course, such numbers as high as 90% are not sensible in reality at all, but we illustrate here how classification results changed in the course of increasing

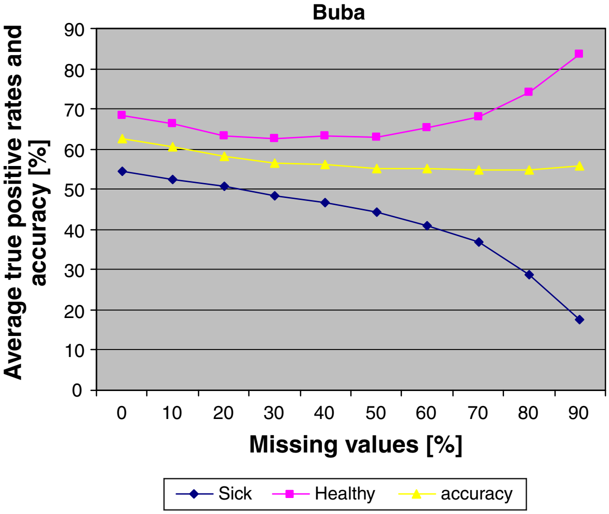

Fig. 1 Average true positive rates and accuracies of the Buba dataset when HEOM was used for the distance calculations of 1-nearest neighbour searching

numbers of missing values. Instead, in the earlier study they tested up to 40% only (Markey et al. 2006).

Originally, the Buba, Haberman and New-thyroid datasets included no missing values. Instead, the Incontinence dataset included approximately 20% missing values and the vertigo dataset around 10 %. In order to simulate complete datasets, when HEOM was used, these missing values were first substituted with either a median or mode depending on the type of variable as mentioned above. The medians and modes were computed class by class in order to restore the two datasets as if they were originally complete. Only after this substitution procedure missing values were randomly selected. For the second test series after the actual imputation the Incontinence and Vertigo datasets were used as such by starting to consider them from the conditions of 20 and 10% missing values, respectively. The imputation was then performed on the basis of modes and medians computed from the whole datasets (not class by class) since it was now our objective to simulate actual classification situations without any prior information about the classes of test cases.

Figures 1, 2, 3, 4 and 5 introduce results computed when HEOM was used to handle distance calculations when missing values were present. Those figures describe both average true positive rates and accuracies. Figure 6 considers results when the imputation approach was employed. Figure 6 only includes accuracies. Here the true positive rates were excluded since their characters essentially resembled those in Figs. 1, 2, 3, 4 and 5.

True positive rate is defined with

$$t_k = \frac{r_k}{n_k} 100\%$$

in which r_k is equal to the number of correctly classified cases in class k and n_k is that of all test cases in class k. Accuracy is given by

$$a = \frac{\sum_{k=1}^{c} r_k}{n} 100\%$$

where c is the number of classes and n all test cases.

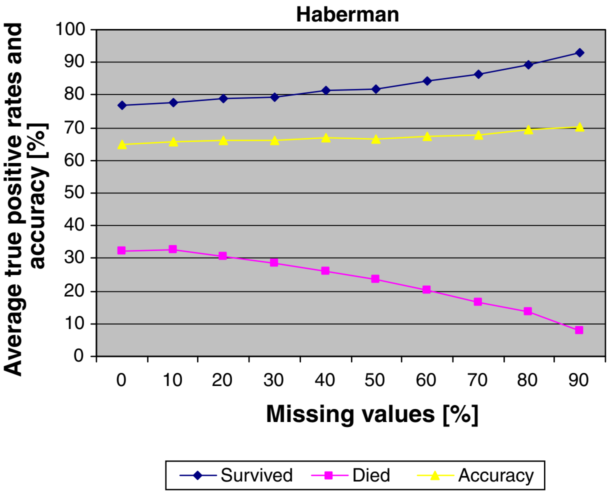

Fig. 2 Average true positive rates and accuracies of the Haberman dataset when HEOM was used for the distance calculations of 1-nearest neighbour searching

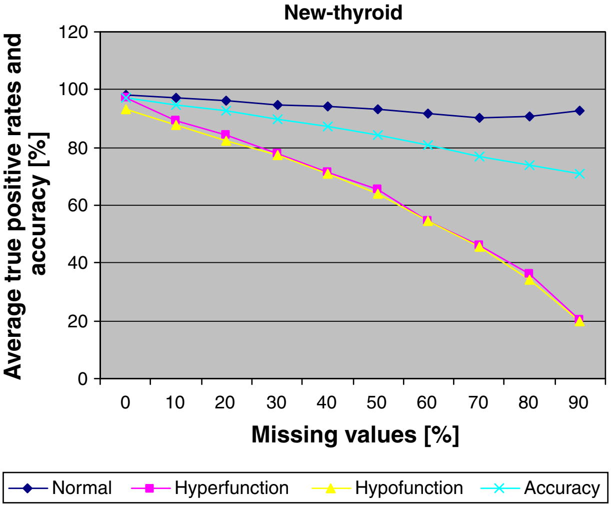

Fig. 3 Average true positive rates and accuracies of the New-thyroid dataset when HEOM was used for the distance calculations of 1-nearest neighbour searching

Standard deviations were computed, but not presented to keep figures clearer. They were mostly small, less than 10% for all subsequent results. Figure 1 shows results given by the Buba dataset. Figure 2 illustrates results computed from the Haberman dataset. Both datasets include two classes. Therefore, the larger class regarding the number of cases in each class begins to predominate with improving true positive rates in 40% missing cases in the Buba dataset and already in 10% of missing cases in the Haberman dataset. This means that the larger class, say the "majority class" obtains slowly improving true positive rates, while the smaller one, "minority" class obtains poorer true positive rates. For the Haberman dataset, the average accuracies computed even increase slightly along with the number of

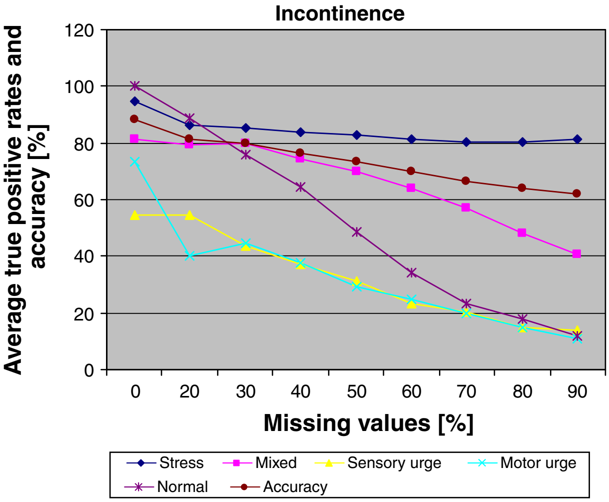

Fig. 4 Average true positive rates and accuracies of the incontinence dataset when HEOM used for the 1-nearest neighbour searching

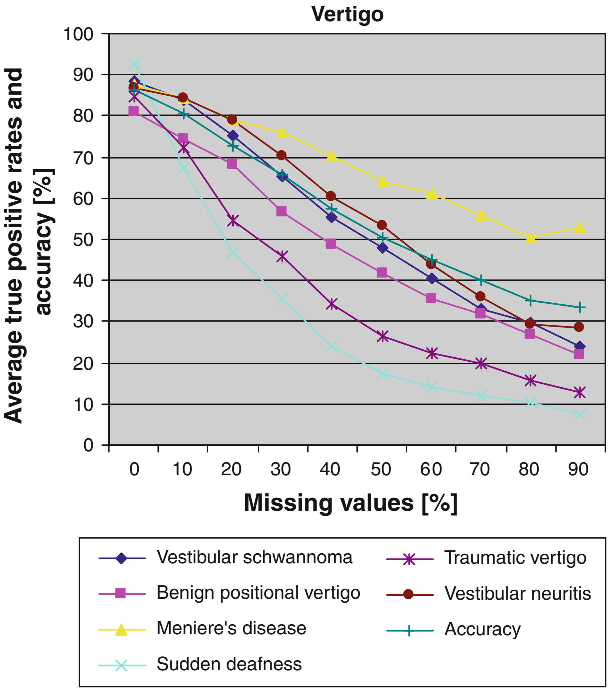

Fig. 5 Average true positive rates and accuracies of the Vertigo dataset when HEOM used for the 1-nearest neighbour searching

increasing missing values. This comes from the property that the majority ‘healthy’ class with 74% (225) from 306 in the Haberman dataset is clearly greater than ‘survivors’ 58% (200) from 345 of the Buba dataset, where this is quite close to 50–50% between the two classes.

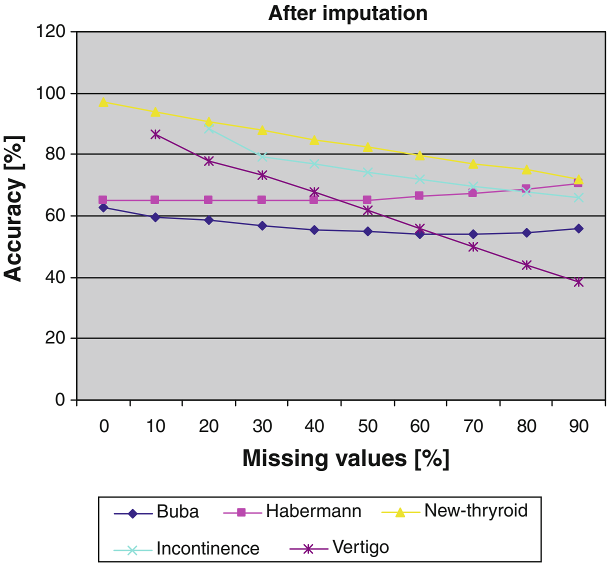

Fig. 6 Average accuracies of all datasets imputed with modes and medians before the 1-nearest neighbour searching

Even the conditions of 30% missing values for the Buba dataset and 40% for the Haberman dataset generated tolerable true positive rates which decreased no more than approximately 6% compared to the situation of no missing value.

Figure 3 indicates the results with the New-thryroid dataset. Now there are three classes and the majority class ‘normal function’ of 70% (150) of the 215 cases is much greater than the two others together. Thus, the true positive rates of the two small minority classes begin to get worse after 10% missing values. When the majority class produced high true positive rates, the average accuracy deteriorates slowly. This situation seems to be typical with imbalanced classes, when there are great differences between the sizes of majority and minority classes. The conditions of 10 or 20% missing values were still tolerable with an approximate 10% decrease of the true positive rates of the minority classes.

Figure 4 contains the results computed with the Incontinence dataset, which includes five classes. The majority class of 61% (323) of the 529 cases again yielded the best true positive rates and the three smallest classes were poor except the class ‘normal’ for 20% of missing values. Nevertheless, the ‘normal’ class was the only that was entirely detected from the other in the situation of no missing value. Thus, it can be seen as an exceptional class. The condition of 20% missing values was acceptable with a maximum drop of 10% in the true positive rates, except for the smallest ‘motor urge’ class. In Fig. 4 the accuracy values slowly get down.

Figure 5 depicts the results obtained from the Vertigo dataset. ‘Menière’s disease’ is the majority class with 39% (313) of the 815 cases. Once again, the majority class produced the best true positive rates along with increasing missing values. Because its part was not so predominant as above, it clearly got worse in the course of increasing missing values. The true positive rates of the two smallest classes of ‘sudden deafness’ and ‘traumatic vertigo’ immediately began to deteriorate after the increase of missing values. The condition of 10% missing values was acceptable for the classes with a maximum drop of approximately 10%, except 15% for the smallest ‘Sudden deafness’ class, and even 20% for the other than the

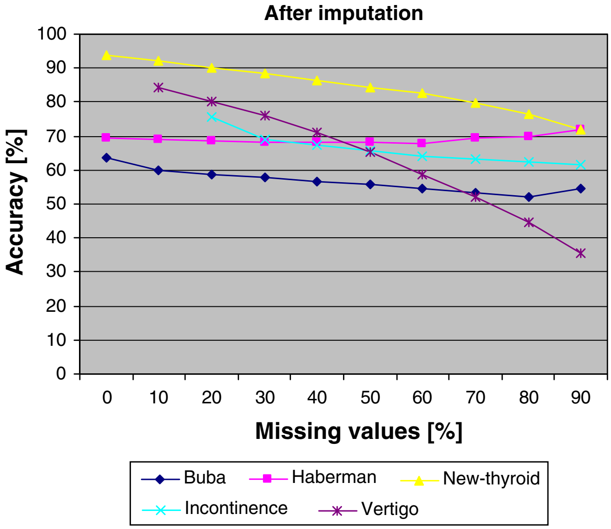

Fig. 7 Average accuracies of all datasets imputed with modes and medians before the 3-mearest neighbour searching

two smallest classes. The accuracy values more rapidly went down than in Fig. 4. There was one class more than in Fig. 4 and the majority class was not as predominant as in Fig. 4.

Next we made another test series imputing missing values with either variable modes or medians. Because originally the Incontinence dataset included approximately 20% of missing values and the Vertigo dataset 10% of missing values, we started from those conditions with them. Otherwise, we randomly marked values as missing as described above and executed 1-nearest neighbour searching 100 times for every test condition with missing values. The average accuracy results are seen in Fig. 6. Again the average accuracies slowly increased after 60% missing values for the Buba dataset and after 30% missing values for the Haberman dataset which is seemingly contradictory, but is caused by the ratios of their two classes. The three other datasets of three, five or six classes behaved more "naturally" while increasing missing values. True positive rates not presented here were clearly alike as those in Figs. 1, 2, 3, 4 and 5. Standard deviations computed were also mostly below 10% for all test conditions with missing values. The average accuracies in Fig. 6 were very similar to those in Figs. 1, 2, 3, 4 and 5 for other than the Vertigo dataset, for which the results in Fig. 6 were 8% better on average. Otherwise, the differences were at most 1–2% larger or less between the two test series.

Next we computed k-nearest neighbour searching ($k = 3$ or 5) to deal with greater k values. This and all subsequent tests were executed after having imputed with medians and modes, similarly to the tests with the results described in Fig. 6. This approach was chosen since the differences between the results were virtually non-existent between HEOM and imputation, except for the largest Vertigo dataset including more classes and variables than the other. Figure 7 shows accuracy results with k equal to 3. The results of k equal to 5 not presented here were almost similar to those of Fig. 7. The results in Fig. 7 indicate mainly small differences compared to Fig. 6. For the Buba dataset, k equal to 3 produced better accuracies, less than approximately 1%. For the Haberman dataset, k equal to 1 aroused 1–4% better accuracies. For the New-thyroid dataset, the situation varied between k equal to 1 and 3 and the differences were 0–3%. For the Vertigo dataset, k equal to 3 yielded

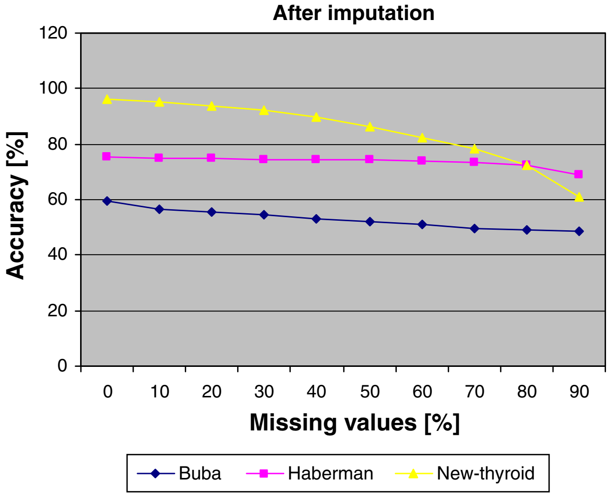

Fig. 8 Average accuracies of all datasets imputed with modes and medians before the naïve Bayesian classification

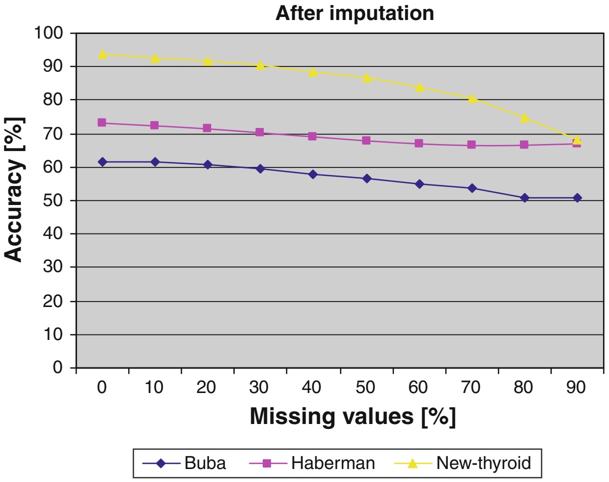

Fig. 9 Average accuracies of all datasets imputed with modes and medians before the linear discriminant analysis

1–3% better accuracies. However, there were significant drops as large as 4–10% for the Incontinence dataset from Figs. 6 to 7. Obviously, the increase of k to 3 began to disturb correct decisions on its three small classes, from which two included so few as 15 and 18 cases. These were the least classes of all five datasets exploited. While comparing accuracies between k equal to 3 and 5, the situations varied between them for the Buba, New-thyroid and Incontinence datasets including differences mostly less than 1%, whereas for Haberman and Vertigo datasets, k equal to 5 caused 1–2% better accuracies.

Next we classified with naïve Bayesian (Fig. 8) and linear discriminant analysis (Fig. 9) techniques after the imputation with medians and modes. The Buba, Haberman and New-thyroid datasets were tested, because the other two aroused frequently such matrices which were not possible to invert. Looking at the accuracies of nearest neighbour searching with k equal to 3 we obtained the following results. Nearest neighbour searching and linear discriminate analysis yielded 2–5% better accuracies than naïve Bayesian technique for the Buba dataset. For the Haberman dataset, naïve Bayesian technique produced 2–4% better than nearest neighbour searching, while linear discriminant was between them. For the Newthyroid dataset, the situation was similar to that of Buba, but for 60% missing values and over naïve Bayesian technique was slightly inferior to two other.

Ultimately, we studied the influence of missing values on classification accuracies when actual missing value distributions were followed variable by variable in each class. Because the Incontinence and Vertigo datasets only included missing values, these datasets were tested here. When the former included originally 19.9% and the latter 13.3% missing values, as to the total numbers of variable values, we applied multiplied numbers of missing values marked randomly. At first, we calculated the numbers of the missing values of all variables in every class. To insert missing values, we followed these distributions in order to simulate their "natural" occurrences. This was done by doubling, tripling etc. their numbers variable by variable in each class. Total numbers of missing values in a class varied from 10 to 36% in the Incontinence dataset and from 9 to 21% in the Vertigo dataset. There could be even more than 50% of missing values for a few individual variables in some classes. Therefore, by doubling the numbers of their missing values, this deleted all original values of such a variable in some class. Similarly, it happened if there were originally 25% of missing values and missing values were inserted three times. After all, these situations were rare, since in most situations there were less than 10% missing values per a variable and a class. In some situations there were no missing value per a variable and a class. Then, of course, no missing value was inserted for such a variable.

We performed the tests by using nearest neighbour searching with k equal to 3 after imputing with medians and modes as above. To simulate real classification circumstances, we computed means and modes over the whole test data, i.e. no classwise knowledge was used for them. Assuming that we would know all medians and modes classwise in a manner of supervised training with the known class labels of a training, our results could be better. The latter is possible in reality when there is a data sample of known class labels available. However, we restricted ourselves to the more difficult condition of unsupervised training data, frequently encountered in real applications, e.g. in respect of medical datasets.

Figure 10 of the Incontinence results shows how the accuracies followed fairly similarly those in Figs. 4, 6 and 7, but the true positive rates differed from those in Fig. 4. These differences were mainly caused by the different missing value distributions. The variation of the true positive values of disease 'Motor urge' obviously comes from the property of this smallest class of all datasets, 15 or 2.8% only from 529 cases. Random influence might appear in this small subset. Using here more and more medians and modes instead of original values occasionally improved the accuracies after the first condition of 20% missing values. The corresponding event was seen for the third least class of 'sensory urge' (6.2% from 529), but rather opposite for the second least of 'normal' (3.4%). Comparing the disease-by-disease true positive rates to those in Fig. 4, we have to remember that, in addition to the different missing value distributions, we also applied HEOM in Fig. 4, but imputation in Fig. 10. Further, k of the former was 1 and that of the latter 3. In Fig. 10 missing values were increased up to 55%, as if a "saturation point", since almost all values still present were then in the variables with no missing values at all at the beginning. Other variables were either completely dead (all values missing) or with rare present values.

Figure 11 shows true positive rates and accuracies of the Vertigo dataset using the "natural" missing value distributions. Both accuracies and disease-by-disease true positive rates are superior in Fig. 11 to those in Figs. 5, 6 and 7. The only exception is the disease 'Sudden deafness', which is very difficult to classify after increasing the number of missing values above approximately 20%. Once again, this is the least class of the current dataset. It incorporates 41 cases (5.0%) only from the whole Vertigo dataset.

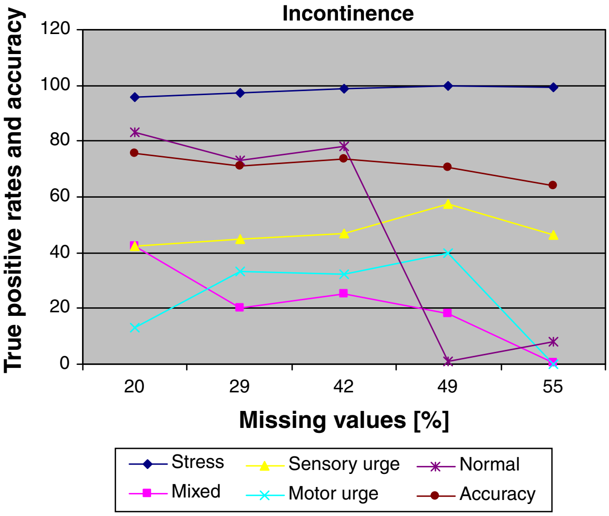

Fig. 10 Average accuracies of the incontinence dataset imputed with modes and medians before the 3-nearest neighbour searching, when missing values were increased randomly along with variable-by-variable “natural” distributions in each class

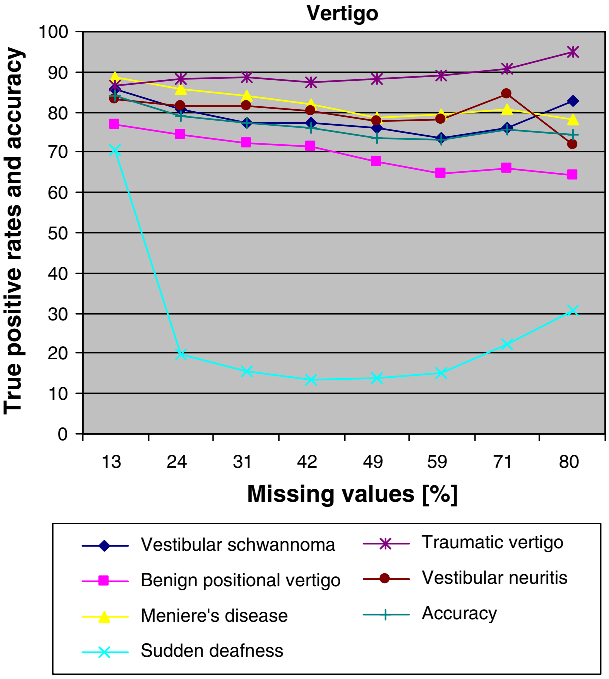

Fig. 11 Average accuracies of the Vertigo dataset imputed with modes and medians before the 3-nearest neighbour searching, when missing values were increased randomly along with variable-by-variable “natural” distributions in each class

5 Discussion

The two-class circumstances associated with the Buba and Haberman datasets in Fig. 1 were slightly surprising because of the gently increasing accuracies, but after all logical and coherent. Let p_1 and p_2 be the probabilities of two-class classification. Since it is always

$$p_1 + p_2 = 1$$

we can name the majority class to be the one with the greater probability assuming that they are not equal. If, for instance, $p_1 >> p_2$ (within interval (0,1)), we can see how the majority class 1 predominates when the number of missing values increases. Under highly missing information classification results start to approach an entirely random guess, i.e. a priori probability given by p_1 that can be estimated with the frequency of the majority class in a two-class dataset. Consequently, when it is equal to $225/306 \approx 0.74$ (cf. the rightmost accuracy in per cent in Fig. 2) in the Haberman dataset, the accuracy approaches this along with the increasing number of the missing values. If there were no data available, the pure guess would yield the majority class with this probability. Generally, for c classes we define a simple baseline value which we call "majority guess"

$$b = \frac{f_{\max}}{\sum_{i=1}^{c} f_i} = \frac{f_{\max}}{n} = p_{\max}$$

in which $f_{\max}$ and $p_{\max}$ are the frequency and a priori probability estimate of the majority class and the sum of all class frequencies f_i is the number n of all cases. In the classification, an acceptable accuracy should exceed b to succeed better than just randomly. In extreme situations such as the Haberman dataset in Fig. 2, where the majority class prevails intensively in an imbalanced class distribution, it is not definite. In this sense Fig. 2 also shows how it is essential to look at the true positive rates of the classes in addition to accuracy values. Of course, several other measures like specificity and positive predictive values not evaluated here are useful.

For the Buba dataset b is $200/345 \approx 0.58$ (cf. the rightmost accuracy in Fig. 1). For the New-thyroid dataset it is $150/215 \approx 0.70$ (Fig. 3). For the Incontinence and Vertigo datasets they are 0.61 (Fig. 4) and 0.38 (Fig. 5). These trends are also seen in the accuracy values in Fig. 6.

When we use HEOM and there are plenty of missing values, the pessimistic way to handle missing values in HEOM increases distances between cases. Such an average distance increases probability to erroneously move a case from a small class to the largest than other way round. Thus, the majority class starts to prevail while increasing missing values. Similarly, using imputations of missing values variable by variable on the basis of the whole data modifies imputed values to more resemble the properties of the majority class than others.

Interestingly, the HEOM and imputation technique were virtually equally efficient for all other datasets than Vertigo for which the imputation technique beat with some 8% better accuracies. The obvious reason is the greatest number 40 of the variables and their types. The number of the variables is 13 in Incontinence and 3, 5 and 6 in other three datasets. Since all Incontinence variables are binary, i.e. their differences between any two cases are either 0 or 1, the dispersion between different distances in HEOM is far smaller than in the Vertigo dataset with 40 variables from which 16 are real and only 14 binary. For the Incontinence dataset there are only integers $\{0, 1, 2, \ldots, 13\}$ possible for the sum of HEOM. Meanwhile, for the Vertigo dataset there are numerous real values possible in interval [0,40] for the sum of HEOM. This makes more probable that there are differences between the results of two techniques for the Vertigo dataset, but not for the Incontinence dataset. In the three other datasets, the numbers of the variables are so small that obviously they did not affect. When there are several variables and more and more missing values in them, HEOM produces greater and greater distances between cases. Thus, HEOM then separates cases from each other, whereas the imputation technique makes cases more similar.

In the beginning of the present research we randomly selected missing values. This can be too simple an assumption in reality. We may assume that they are not always missing randomly, but some values are deliberately ignored. For instance, a physician may have a strong hypothesis about the diagnosis and, consequently, he or she sees some variables (e.g. laboratory tests) unnecessary for some disease. That is why several values may be absent on purpose. In a way, this might result in a stronger endurance of missing values than those appeared here in Figs. 1, 2, 3, 4, 5, 6, 7, 8 and 9. This possibility was supported by the results in Figs. 10 and 11, where missing values were increased along with their actual distributions in the Incontinence and Vertigo datasets. The class 'Stress' of Incontinence in Fig. 10 clearly took advantage of its majority property. The small classes 'Sensory urge', 'Motor urge' and 'Normal' lost most. In addition, the next largest 'Mixed' lost expressly. After all, it included 26% only from the whole dataset, less than a half from 'Stress' with 61%. In Fig. 11 'Sudden deafness', the least class, rapidly dropped while increasing missing values. Instead, the next least class, 'traumatic vertigo' did not suffer from increasing missing values, but even obtained the best results of all classes. The reason was that it included no missing values in its class for its most important variable (the presence of head trauma), when typically all cases in this class have had a head trauma, but not in the other classes. Apart from the least class, the Vertigo dataset was surprisingly tolerant of increasing missing values. This supports the rationale that the original (actual) missing values (13.3%) of the dataset were not missing at random, but the physicians were mostly correctly left them out understanding some variables to be only slightly useful to investigate for some of their patient cases.

In the future we are going to extend our research to other medical datasets. It would be useful to study larger datasets with more variables, cases, classes and class distributions. Further, more classification methods could be studied to see whether they tolerate more or less missing values than nearest neighbour searching, linear discriminant analysis and naïve Bayesian classification.

6 Summary

Overall the two ways to consider missing values, i.e. the pessimistic way in the heterogeneous Euclidean-overlap metric (HEOM) and imputation with variable modes and medians, seem to be virtually equally good means for a small dataset with few classes. The only exception in our datasets was Vertigo including more variables and more classes than the other. The imputation technique was then better.

If there are only two classes and the majority class is not clearly greater than the minority class, these simple circumstances allow fairly many missing values, even 30–40%, simply because the situation is near a random guess. The more imbalanced class distribution, the more sensitive situation to missing values is found. At their largest, the situations with 20% missing values of all were then tolerable among the datasets tested. Also, the more classes, the more sensitive to missing values is encountered. This is especially true if the class distribution is rather imbalanced at the same time. In our tests the conditions with 10% as missing values were then an actual maximum. These results indicate that classifications under missing value circumstances have to be considered carefully without lapsing into the substitution of too frequent missing values. On the other hand, it is also possible that sometimes a dataset may be fairly tolerant of missing values, since at least for medical datasets missing values do not often miss randomly but after a purposeful selection. We restricted ourselves chiefly to the straightforward and heuristic nearest neighbour searching. Some more complicated classification methods might suffer differently from missing values.

Acknowledgments The second author had collected the Incontinence dataset at the Department of Obstetrics and Gynaecology in the Kuopio University Hospital, Finland. The Vertigo dataset was from the Department of Otorhinolaryngology, Helsinki University Central Hospital, Finland. The authors acknowledge their staffs.

References

Aha DW, Kibler D, Albert MK (1991) Instance-based learning algorithms. Mach Learn 6:37–66

Blake CL, Merz CJ (1998) UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine. http://www.ics.uci.edu/~mlearn/MLRepository.html/

Chowdhury S, Bodemar G, Haug P, Bapic A, Wigertz O (1991) Methods for knowledge extraction from a clinical database on liver diseases. Comput Biomed Res 24:530–548

Fortes I, Mora-López L, Morales R, Triguere F (2006) Inductive learning models with missing values. Math Comp Modell 44:790–806

Laurikkala J, Juhola M (1998) Genetics-based machine learning system to discover diagnostic rules for female urinary incontinence. Comput Meth Prog Biomed 55:217–228

Laurikkala J, Juhola M, Lammi S, Penttinen J, Aukee P (2001) Analysis of the imputed female urinary incontinence data for the evaluation of expert system parameters. Comp Biol Med 31:239–257

Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York

Markey MK, Tourassi GD, Margolis M, DeLong DM (2006) Impact of missing data in evaluating artificial neural networks trained on complete data. Comp Biol Med 36:516–525

Mykkänen J, Juhola M, Ruotsalainen U (2000) Extracting VOIs from brain PET images. Int J Med Inf 58(59):59–69

Pesonen E, Eskelinen M, Juhola M (1998) Treatment of missing data values in a neural network based decision support system for acute abdominal pain. Artif Intell Med 13:139–146

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

Viikki K, Kentala E, Juhola M, Pyykkö I (1999) Decision tree induction in the diagnosis of otoneurological diseases. Med Inf Internet Med 24:277–289

Wasito I, Mirkin B (2005) Nearest neighbour approach in the least-squares data imputation algorithms. Inf Sci 169:1–25

Wilson DR, Martinez TR (1997) Improved heterogeneous distance functions. J Artif Intell Res 6:1–34

Witten IH, Frank E (2000) Data mining, practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco