Handling Missing Attribute Values in Preterm Birth Data Sets

Jerzy W. Grzymala-Busse[1], Linda K. Goodwin[2], Witold J. Grzymala-Busse[3], and Xinqun Zheng[4]

[1] Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA
[2] Institute of Computer Science, Polish Academy of Sciences, 01-237 Warsaw, Poland
Jerzy@ku.edu, http://lightning.eecs.ku.edu/index.html
[3] Nursing Informatics Program, Duke University, Durham, NC 27710, USA
Linda.Goodwin@duke.edu
[4] Filterlogix, Lawrence, KS 66049, USA
WBusse@FilterLogix.com
[5] PC Sprint, Overland Park, KS 66211, USA
Xinqun.Zheng@mail.sprint.com

Abstract. The objective of our research was to find the best approach to handle missing attribute values in data sets describing preterm birth provided by the Duke University. Five strategies were used for filling in missing attribute values, based on most common values and closest fit for symbolic attributes, averages for numerical attributes, and a special approach to induce only certain rules from specified information using the MLEM2 approach. The final conclusion is that the best strategy was to use the global most common method for symbolic attributes and the global average method for numerical attributes.

1 Introduction

Predicting preterm birth risk among pregnant women is a difficult problem. Diagnosis of preterm birth is attributed with a positive predictive value (the ratio of all true positives to the sum of all true positives and false positives) only between 17 and 38% [7].

The main objective of our research was to find the best approach to handling missing attribute values in data sets describing preterm birth. These data, collected at the Duke University, were affected by vast quantity of missing attribute values. Additionally, in spite of the fact that many attributes were numerical, these data sets were inconsistent, another complication for data mining.

Additionally, the best approach to missing attribute values must be selected taking into account that the main criterion of quality is not the smallest error rate but the sum of sensitivity (conditional probability of diagnosis of preterm birth) and sensitivity (conditional probability of diagnosis of fullterm birth). In

D. Ślęzak et al. (Eds.): RSFDGrC 2005, LNAI 3642, pp. 342–351, 2005.

order to increase sensitivity, an additional technique of changing rule strength was applied [6]. Another important criterion of rule quality is the area under the curve for the ROC graph.

2 Missing Attribute Values

In this paper we will discuss only methods dealing with incomplete data sets (with missing attribute values) based on conversion of incomplete data sets into complete data sets, without missing attribute values. Such a conversion is conducted before the main process of rule induction, therefore it is a kind of preprocessing.

2.1 Global Most Common Attribute Value for Symbolic Attributes, and Global Average Value for Numerical Attributes (GMC-GA)

This method is one of the simplest methods among the methods to deal with missing attribute values. For symbolic attributes, every missing attribute value should be replaced by the most common attribute value; for numerical attributes, every missing value should be replaced by the average of all values of the corresponding attribute.

2.2 Concept Most Common Attribute Value for Symbolic Attributes, and Concept Average Value for Numerical Attributes (CMC-CA)

This method may be considered as the method from Subsection 2.1 restricted to concepts. A concept is a subset of the set of all cases with the same outcome. In preterm birth data sets there were two concepts, describing preterm and fullterm birth. In this method, for symbolic attributes, every missing attribute value should be replaced by the most common attribute value that occurs for the same concept; for numerical attributes, every missing values should be replaced by the average of all values of the attributed, restricted to the same concept.

2.3 Concept Closest Fit (CCF)

The closest fit algorithm [4] for missing attribute values is based on replacing a missing attribute value with an existing value of the same attribute from another case that resembles as much as possible the case with missing attribute values. When searching for the closest fit case, we need to compare two vectors of attribute values of the given case with missing attribute values and of a searched case.

During the search, for each case a proximity measure is computed, the case for which the proximity measure is the smallest is the closest fitting case that is used to determine the missing attribute values. The proximity measure between two cases x and y is the Manhattan distance between x and y, i.e.,

$$distance(x, y) = \sum_{i=1}^{n} distance(x_i, y_i),$$

where

$$distance(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \ , \\ 1 & \text{if } x \text{ and } y \text{ are symbolic and } x_i \neq y_i, \\ & \text{or } x_i = ? \text{ or } y_i = ?, \\ \frac{|x_i - y_i|}{r} & \text{if } x_i \text{ and } y_i \text{ are numbers and } x_i \neq y_i \ , \end{cases}$$

where r is the difference between the maximum and minimum of the known values of the numerical attribute with a missing value. If there is a tie for two cases with the same distance, a kind of heuristics is necessary, for example, select the first case. In general, using the global closest fit method may result in data sets in which some missing attribute values are not replaced by known values. Additional iterations of using this method may reduce the number of missing attribute values, but may not end up with all missing attribute values being replaced by known attribute values.

3 Duke Data Sets

The preterm birth data were collected at the Duke University Medical Center. This data set includes a sample of 19,970 ethnically diverse women and includes 1,229 variables. The data set was partitioned into two parts: training (with 14,977 cases) and testing (with 4,993 cases). Three mutually disjoint subsets of the set of all 1,229 attributes were selected, the first set contains 52 attributes, the second 54 attributes and the third subset contains seven attributes; the new data sets were named Duke-1, Duke-2, and Duke-3, respectively. The Duke-1 set contains laboratory test results. The Duke-2 test represents the most essential remaining attributes that, according to experts, should be used in diagnosis of preterm birth. Duke-3 represents demographic information about pregnant women. All the three data sets are large, have many missing attribute values, are unbalanced, many attributes are numerical, and the data sets are inconsistent. Tables 1 and 2 outline the basic characteristics of these three data sets.

Table 1. Duke training data sets

	Duke-1	Duke-2	Duke-3
Number of cases	14,997	14,997	14,997
Number of attributes	52	54	7
Number of concepts	2	2	2
Consistency level	42.18%	47.61%	95.95%
Number of cases in the basic class	3,116	3,116	3,069
Number of cases in the complementary class	11,861	11,861	11,908
Number of missing atribute values	503,743	291,338	4,703

Table 2. Duke testing data sets

	Duke-1	Duke-2	Duke-3
Number of cases	4,993	4,993	4,993
Number of attributes	52	54	7
Number of concepts	2	2	2
Consistency level	42.34%	52.29%	98.52%
Number of cases in the basic class	1,010	1,010	1,057
Number of cases in the complementary class	3,983	3,983	3,936
Number of missing attribute values	168,957	97,455	1,618

4 Data Mining Tools

In our experiments, for rule induction the algorithm LEM2 (Learning from Examples Module, version 2) was used [2]. LEM2 is a component of the LERS (Learning from Examples based on Rough Sets) data mining system. Additionally, a modified version of LEM2, called MLEM2, was also used for some experiments [3]. The classification system of LERS is a modification of the bucket brigade algorithm. The decision to which concept a case belongs is made on the basis of three factors: strength, specificity, and support. They are defined as follows: *Strength* is the total number of cases correctly classified by the rule during training. *Specificity* is the total number of attribute-value pairs on the left-hand side of the rule The third factor, *support*, is defined as the sum of scores of all matching rules from the concept, where the score of the rule is the product of its strength and specificity. The concept for which the support is the largest is the winner and the case is classified as being a member of that concept.

5 Criteria Used to Measure the Rule Set Quality

Several criteria were used to measure the rule set quality in our experiments: error rate, sensitivity and specificity, and the area under curve (AUC) of the receiver operating Characteristic (ROC) [8]. For unbalanced data sets, error rate is not a good indicator for rule set quality. $Sensitivity + Specificity - 1$ is a better indicator as well as the Area Under Curve of Receiver Operating Characteristic.

5.1 Error Rate

In medical diagnosis, the objective is not to achieve a small error rate. Diagnosticians are interested mostly in correctly diagnosing the cases that are affected by disease. Moreover, frequently medical data sets are unbalanced: one class is represented by the majority of cases while the other class is represented by the minority. Unfortunately, in medical data the smaller class—as a rule—is more

important. We will call this class basic, and the other class complementary. Consequently, the error rate in the original rule sets is not a good indicator of rule set quality [6].

5.2 Sensitivity and Specificity

The set of all correctly classified (preterm) cases from the basic concept are called true-positives, incorrectly classified basic cases (i.e., classified as fullterm) are called false-negatives, correctly classified complementary (fullterm) cases are called true-negatives, and incorrectly classified complementary (fullterm) cases are called false-positives.

Sensitivity is the conditional probability of true-positives given basic concept, i.e., the ratio of the number of true-positives to the sum of the number of true-positives and false-negatives. It will be denoted by $P(TP)$. Specificity is the conditional probability of true-negatives given complementary concept, i.e., the ratio of the number of true-negatives to the sum of the number of true-negatives and false-positives. It will be denoted by $P(TN)$. Similarly, the conditional probability of false-negatives, given actual preterm, and equal to $1 - P(TP)$, will be denoted by $P(FN)$ and the conditional probability of false-positives, given actual fullterm, and equal to $1 - P(TN)$, will be denoted by $P(FP)$.

In Duke's prenatal training data, only 20.7% of the cases represent the basic concept, preterm birth. During rule induction, the average of all rule strengths for the bigger concept is also greater than the average of all rule strengths for the more important but smaller basic concept. During classification of unseen cases, rules matching a case and voting for the basic concept are outvoted by rules voting for the bigger, complementary concept. Thus the sensitivity is poor and the resulting classification system would be rejected by diagnosticians.

Therefore it is necessary to increase sensitivity by increasing the average rule strength for the basic concept. In our research we selected the optimal rule set by multiplying the rule strength for all rules describing the basic concept by

Table 3. Duke-1, only certain rules

	GMC-GA	CMC-CA	CCF-CMC-CA	CCF-CMC	CCF-MLEM2
Initial error rate	21.29%	20.25%	64.65%	20.89%	20.39%
Critical error rate	40.48%	N/A	N/A	61.39%	56.68%
MAX $P(TP) - P(FP)$	0.156	0	-0.469	0.0782	0.1062
MIN $P(TP) - P(FP)$	-0.009	-0.122	-0.4845	-0.0895	-0.0588
Critical rule strength multiplier	7.7	N/A	N/A	13.38	6.5
AUC	0.5618	0.4563	0.2602	0.4878	0.5197

Table 4. Duke-2, only certain rules

	GMC-GA	CMC-CA	CCF-CMC-CA	CCF-CMC	CCF-MLEM2
Initial error rate	21.37%	20.23%	21.91%	20.83%	21.41%
Critical error rate	51.79%	N/A	N/A	41.1%	50.47%
MAX $P(TP) - P(FP)$	0.1224	0.0026	-0.0025	0.057	0.1419
MIN $P(TP) - P(FP)$	0.0007	-0.0028	-0.0166	-0.0813	-0.0109
Critical rule strength multiplier	6.6	N/A	N/A	12.07	5
AUC	0.5505	0.5013	0.4952	0.496	0.5624

the same real number called a strength multiplier. In general, the sensitivity increases with the increase of the strength multiplier. At the same time, specificity decreases. An obvious criterion for the choice of the optimal value of the strength multiplier is the maximum of the difference between the relative frequency of true positives, represented by *Sensitivity*, and the relative frequency of false positives, represented by *Specificity − 1*. Thus we wish to maximize

$$Sensitivity + Specificity - 1 = P(TP) - P(FP)$$

This criterion is based on an analysis presented by Bairagi and Suchindran [1]. For each rule set, there exists some value of the strength multiplier, called *critical* (or *optimal*), for which the values of $P(TP) - P(FP)$ is maximum. The total error rate, corresponding to the rule strength multiplier equal to one, is called *initial*; while the total arror rate, corresponding to the critical strength multiplier, is called *critical*.

5.3 The Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) Graph

The ROC graph is a plot of sensitivity versus one minus specificity. The major diagonal, a line that goes through (0, 0) and (1, 1), represents a situation in which the hit and false-alarm are equal. It corresponds to making a random diagnosis. Thus the ROC curve should be located above the main diagonal, the further from the diagonal the better [8]. The bigger the AUC value, the better the quality of the rule set. Apparently, $AUC = 0.5$ corresponds to random diagnosis. So, $AUC > 0.5$ means the result is better than the random diagnosis, and $AUC < 0.5$ means the result is worse than the random diagnosis.

6 Experiments

First, for the three Duke data sets, missing attribute values were replaced using the five methods. The first two methods were GMC-GA and CMC-CA. Since

Table 5. Duke-3, only certain rules

	GMC-GA	CMC-CA	CCF-CMC-CA	CCF-CMC
Initial error rate	22.33%	22.37%	22.55%	22.63%
Critical error rate	48.65%	47.03%	47.45%	50.09%
MAX $P(TP) - P(FP)$	0.1524	0.1578	0.1608	0.1473
MIN $P(TP) - P(FP)$	0.0102	0.0124	0.0122	0.0108
Critical rule strength multiplier	12	11	10	10
AUC	0.5787	0.5888	0.5854	0.5821

Table 6. Only possible rules

	Duke-1		Duke-2		Duke-3	
	GMC-GA	CCF-CMC	GMC-GA	CCF-CMC	GMC-GA	CCF-CMC
Initial error rate	21.95%	20.81%	21.53%	20.85%	23.79%	23.91%
Critical error rate	56.19%	43.98%	53.74%	59.80%	34.15%	31.32
MAX $P(TP) - P(FP)$	0.0894	0.1427	0.0818	0.0522	0.1412	0.1383
MIN $P(TP) - P(FP)$	-0.0437	-0.2114	0.0046	-0.091	0.0193	0.0157
Critical rule strength multiplier	4	6.8	2.1	12.28	10	8
AUC	0.5173	0.5528	0.5383	0.49	0.5707	0.5714

the missing attribute value rates were so high, applying the concept closest fit algorithm (CCF) could not fill in all the missing attribute values in these three data sets. So, the concept most common method for symbolic attributes and the concept average value method for numerical attributes (CMC-CA), and concept most common for both symbolic and numerical attributes method (CMC) were used respectively followed by the method of concept closest fit. For the same reason, the MLEM2 algorithm for Duke-1 and Duke-2 was tested after the concept closest fit algorithm (CCF) was applied.

To reduce the error rate during classification a very special discretization method for Duke-1 and Duke-2 was used. First, in the training data set, for any numerical attribute, values were sorted. Every value v was replaced by the interval $[v, w)$, where w was the next larger value than v in the sorted list. This discretization method was selected because the original data sets, with numerical attributes, were inconsistent.

Table 7. First certain rules, then possible rules

	Duke-1		Duke-2		Duke-3	
	GMC-GA	CCF-CMC	GMC-GA	CCF-CMC	GMC-GA	CCF-CMC
Initial error rate	21.89%	21.03%	21.65%	20.91%	23.91%	24.03%
Critical error rate	41.0%	59.74%	51.67%	41.04%	34.97%	38.33%
MAX						
$P(TP) - P(FP)$	0.155	0.0841	0.1135	0.0533	0.1329	0.1823
MIN						
$P(TP) - P(FP)$	-0.0099	-0.0837	0.002	-0.085	0.0157	0.0142
Critical rule						
strength multiplier	7.7	13.37	6.6	12.07	13	16
AUC	0.562	0.4929	0.5454	0.4929	0.5623	0.5029

In the experiments, four combinations of using rule sets were applied: *using only certain rules*, *using only possible rules*, *using certain rules first then possible rules if necessary*, and *using both certain and possible rules*. The option *complete matching, then partial matching if necessary* is better than the option *using both complete matching and partial matching* [5], so only that first option was used.

For training data sets Duke-1 and Duke-2, the consistency levels were 100% after replacing missing attribute values by methods CMC-CA and by CCF-CMC-CA, so no possible rules were induced. We used MLEM2 only to induce certain rules. Thus in Tables 6–8, only two methods are listed: GMC-GA and CCF-CMC.

From Tables 4 and 5 it is clear that by using methods CMC-CA and CCF-CMC-CA for Duke-1 and Duke-2 the worst results were obtained. Comparing CCF-MLEM2 and CCF-CMC (Tables 3 and 4) based on the $P(TP) - P(FP)$, we can see that the CCF-MLEM2 method provided slightly better results.

Comparison of the four strategies to deal with certain and possible rules was conducted for two methods: GMC-GA and CCF-CMC. The GMC-GA method was the simplest method of the five methods tested and this method produced better results than CCF-CMC (based on the value of $P(TP) - P(FP)$ and AUC). This can be verified by the Wilcoxon matched-pairs signed rank test (5% significance level).

For Duke-3, the four methods GMC-GA, CMC-CA, CCF-CMC-CA, CCF-CMC produced roughly the same results in each classification strategy, see Tables 5–8. The explanation of this result may be that the attributes with missing values were not critical attributes so that any filling in missing values used before rule induction may not affect the quality of rule set greatly.

In order to make the best use of certain and possible rule sets induced by LEM2 from inconsistent data, four different strategies of classification were tested in the experiments. From experiments on Duke-3, see Tables 5–8, it could be seen that using only certain rules provided the biggest value of $P(TP) - P(FP)$ among

Table 8. Union of certain and possible rules

	Duke-1		Duke-2		Duke-3	
	GMC-GA	CCF-CMC	GMC-GA	CCF-CMC	GMC-GA	CCF-CMC
Initial error rate	21.79%	20.95%	21.65%	20.83%	23.47%	23.89%
Critical error rate	53.23%	49.35%	43.44%	41.78%	31.18%	30.64%
MAX $P(TP) - P(FP)$	0.0999	0.1175	0.1532	0.0525	0.1456	0.1366
MIN $P(TP) - P(FP)$	-0.0263	-0.0467	0.0073	-0.0906	0.0227	0.0139
Critical rule strength multiplier	4.3	8.21	2.3	12.17	8	8
AUC	0.5305	0.5304	0.5681	0.4903	0.5707	0.5704

the four strategies based on each of the four filling in missing attribute value methods: GMC-GA, CMC-CA, CCF-CMC-CA, and CCF-CMC. This shows that for low consistency level data sets, certain rules are more important than possible rules.

7 Conclusions

Among the five different filling in missing values methods tested, our results show that for DukeŠs data, GMC-GA provided the best results. This is a result of the poor quality DukeŠs data sets, where the missing rate is very high for many numerical attribute values. For the same reason, applying CMC-CA directly or followed by CCF for DukeŠs data sets, provides worse results.

MLEM2 usually induces fewer rules than other rule induction methods. But it did not produce good results for data sets that have low consistency levels. However, for data sets with high consistency levels, MLEM2 induced high quality rule sets.

By comparing the four strategies of classification methods, the only conclusion is that for low consistency level data sets, certain rules are better than possible rules. On the other hand, for high consistency level data sets, there is no one single best strategy.

References

1. Bairagi, R. and Suchindran C.M.: An estimator of the cutoff point maximizing sum of sensitivity and specificity. Sankhya, Series B, Indian Journal of Statistics 51 (1989) 263–269.
2. Grzymala-Busse, J. W.: LERS—A system for learning from examples based on rough sets. In Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory. Slowinski, R. (ed.), Kluwer Academic Publishers, 1992, 3–18.

3. Grzymala-Busse, J.W.:. MLEM2: A new algorithm for rule induction from imperfect data. Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2002, July 1–5, Annecy, France, 243–250.
4. Grzymala-Busse, J. W., Grzymala-Busse, W. J. and Goodwin, L. K.: A closest fit approach to missing attribute values in preterm birth data. Proc. of the Seventh Int. Workshop on Rough Sets, Fuzzy Sets, Data Mining and Granular-Soft Computing (RSFDGrC'99), Ube, Yamaguchi, Japan, November 8Ű10, 1999. Lecture Notes in Artificial Intelligence, No. 1711, Springer Verlag, 1999, 405–413.
5. Grzymala-Busse, J.W. and Zou X.: Classification strategies using certain and possible rules. Proc. of the First International Conference on Rough Sets and Current Trends in Computing, Warsaw, Poland, June 22–26, 1998. Lecture Notes in Artificial Intelligence, No. 1424, Springer Verlag, 1998, 37–44.
6. Grzymala-Busse, J. W., Goodwin, L.K., and Zhang, X.: Increasing sensitivity of preterm birth by changing rule strengths. Proceedings of the 8th Workshop on Intelligent Information Systems (IIS'99), Ustronie, Poland, June 14–18, 1999, 127–136.
7. McLean, M., Walters, W. A. and Smith, R.:. 1993. Prediction and early diagnosis of preterm labor: a critical review. Obstetrical & Gynecological Survey 48 (1993) 209–225.
8. Swets, J.A. and Pickett, R.M.: Evaluation of Diagnostic Systems. Methods from Signal Detection Theory. Academic Press, 1982.