A Closest Fit Approach to Missing Attribute Values in Preterm Birth Data

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

[1] Department of Electrical Engineering and Computer Science,
University of Kansas, Lawrence, KS 66045, USA
[2] RS Systems, Inc., Lawrence, KS 66047, USA
[3] Department of Information Services and the School of Nursing,
Duke University, Durham, NC 27710, USA

ABSTRACT: In real-life data, in general, many attribute values are missing. Therefore, rule induction requires preprocessing, where missing attribute values are replaced by appropriate values. The rule induction method used in our research is based on rough set theory.

In this paper we present our results on a new approach to missing attribute values called a closest fit. The main idea of the closest fit is based on searching through the set of all cases, considered as vectors of attribute values, for a case that is the most similar to the given case with missing attribute values. There are two possible ways to look for the closest case: we may restrict our attention to the given concept or to the set of all cases. These methods are compared with a special case of the closest fit principle: replacing missing attribute values by the most common value from the concept. All algorithms were implemented in system OOMIS. Our experiments were performed on preterm birth data sets collected at the Duke University Medical Center.

KEYWORDS: Missing attribute values, closest fit, data mining, rule induction, classification of unseen cases, system OOMIS, rough set theory.

1 Introduction

Recently data mining, i.e., discovering knowledge from raw data, is receiving a lot of attention. Such data are, as a rule, imperfect. In this paper our main focus is on missing attribute values, a special kind of imperfection. Another form of imperfection is inconsistency—the data set may contain conflicting cases (examples), having the same values of all attributes yet belonging to different concepts (classes).

Knowledge considered in this paper is expressed in the form of rules, also called production rules. Rules are induced from given input data sets by algorithms based on rough set theory. For each concept lower and upper approximations are computed, as defined in rough set theory [4, 6, 12, 13].

Often in real-life data some attribute values are *missing* (or *unknown*). There are many approaches to handle missing attribute values [3, 5, 7]. In this paper we will discuss an approach based on the closest fit idea. The closest fit algorithm for missing attribute values is based on replacing a missing attribute value by existing values of the same attribute in another case that resembles as much as possible the

N. Zhong, A. Skowron, S. Ohsuga (Eds.): RSFDGrC'99, LNAI 1711, pp. 405-413, 1999.

case with the missing attribute values. In 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.

There are many possible variations of the idea of the closest fit. First, for a given case with a missing attribute value, we may look for the closest fitting cases within the same concept, as defined by the case with missing attribute value, or in all concepts, i.e., among all cases. The former algorithm is called *concept closest fit*, the latter is called *global closest fit*.

Secondly, we may look at the closest fitting case that has all the same values, including missing attribute values, as the case with a missing attribute value, or we may restrict the search to cases with no missing attribute values. In other words, the search is performed on cases with missing attribute values or among cases without missing attribute values.

During the search, the entire training set is scanned, for each case a proximity measure is computed, the case for which the proximity measure is the largest is the closest fitting case that is used to determine the missing attribute values. The proximity measure between two cases e and e' is the Manhattan distance between e and e', i.e.,

$$\sum_{i=1}^{n} \text{distance } (e_i, e_i'),$$

where

$$\text{distance } (e_i, e_i') = \begin{cases} 0 & \text{if } e_i \text{ and } e_i' \text{ are symbolic and } e_i \neq e_i', \\ 1 & \text{if } e_i = e_i', \\ 1 - \dfrac{|e_i - e_i'|}{|a_i - b_i|} & \text{if } e_i \text{ and } e_i' \text{ are numbers and } e_i \neq e_i', \end{cases}$$

where a_i is the maximum of values of A_i, b_i is the minimum of values of A_i, and A_i is an attribute.

In a special case of the closest fit algorithm, called the most common value algorithm, instead of comparing entire vectors of attribute values, the search is reduced to just one attribute, the attribute for which the case has a missing value. The missing value is replaced by the most frequent value within the same concept to which belongs the case with a missing attribute value.

2 Rule Induction and Classification of Unseen Cases

In our experiments we used LERS (Learning from Examples based on Rough Set theory) for rule induction. LERS has four options for rule induction; only one, called LEM2 [4, 6] was used for our experiments. Rules induced from the lower approximation of the class *certainly* describe the class, so they are called *certain*. On the other hand, rules induced from the upper approximation of the class describe only *possibly* (or *plausibly*) cases, so they are called *possible* [8]. Examples of other data mining systems based on rough sets are presented in [14, 16].

For classification of unseen cases system LERS uses a modified "bucket brigade

algorithm" [2, 10]. The decision to which class a case belongs is made on the basis of two parameters: strength and support. They are defined as follows: *Strength* is the total number of cases correctly classified by the rule during training. The second parameter, *support*, is defined as the sum of scores of all matching rules from the class. The class C for which the support, i.e., the value of the following expression

$$\sum_{\text{matching rules } R \text{ describing } C} \text{Strength}(R)$$

is the largest is a winner and the case is classified as being a member of C. The above scheme reminds non-democratic voting in which voters vote with their strengths.

If a case is not completely matched by any rule, some classification systems use *partial matching*. During partial matching, system AQ15 uses the probabilistic sum of all measures of fit for rules [11]. Another approach to partial matching is presented in [14]. Holland *et al*. [10] do not consider partial matching as a viable alternative of complete matching and thus rely on a default hierarchy instead. In LERS partial matching does not rely on the input of the user. If complete matching is impossible, all partially matching rules are identified. These are rules with at least one attribute-value pair matching the corresponding attribute-value pair of a case.

For any partially matching rule R, the additional factor, called *Matching_factor* (R), is computed. Matching_factor is defined as the ratio of the number of matched attribute-value pairs of a rule with a case to the total number of attribute-value pairs of the rule. In partial matching, the class C for which the value of the following expression

$$\sum_{\text{partially matching rules } R \text{ describing } C} \text{Matching\_factor}(R) * \text{Strength}(R)$$

is the largest is the winner and the case is classified as being a member of C.

During classification of unseen (testing) cases with missing attribute values, missing attribute values do not participate in any attempt to match a rule during complete or partial matching. A case can match rules using only actual attribute values.

3 Description of Data Sets and Experiments

Data sets used for our experiments come from the Duke University Medical Center. First, a large data set, with 1,229 attributes and 19,970 cases was partitioned into two parts: training (with 14,977 cases) and testing (with 4,993 cases). We selected two mutually disjoint subsets of the set of all 1,229 attributes, the first set containing 52 attributes and the second with 54 attributes and called the new data sets Duke-1 and Duke-2, respectively. The Duke-1 data 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. Both data sets were unbalanced because only 3,103 cases were preterm, all remaining 11,874 cases were fullterm.

Table 1. Missing attribute values

	Number of missing attribute values in data sets processed by		
	Global closest fit	Concept closest fit	Most common value
Duke-1	1,1641	505,329	0
Duke-2	615	1,449	0

Similarly, in the testing data set, there were only 1,023 preterm cases while the number of fullterm cases was 3,970.

Both data sets, Duke-1 and Duke-2, have many missing attribute values (Duke-1 has 505,329 missing attribute values, i.e., 64.9% of the total number of attribute values; Duke-2 has 291,796 missing attribute values, i.e., 36.1% of the total number of attribute values).

First, missing attribute values were replaced by actual values. Both data sets were processed by the previously described five algorithms of the OOMIS system: global closest fit and concept closest fit, among all cases with and without missing attribute values, and most common value.

Since the number of missing attribute values in Duke-1 or Duke-2 is so large, we were successful in using only three algorithms. The version of looking for the closest fit *among all cases without missing attribute values* returned the unchanged, original data sets. Therefore, in the sequel we will use names *global closest fit* and *concept closest fit* for algorithms that search among all cases with missing attribute values. For Duke-1 the concept closest fit algorithm was too restrictive: All missing attribute values were unchanged, so we ignored the Duke-1 data set processed by the concept closest fit algorithm. Moreover, global closest fit or concept closest fit algorithms returned data sets with only reduced number of missing attribute values. The results are presented in Table 1.

Since using both closest fit options result in some remaining missing attribute values, for the output files the option most common value was used to replace all remaining missing attribute values by the actual attribute values. Thus, finally we

Table 2. Training data sets

		Global closest fit	Concept closest fit	Most common value
Duke-1	Number of conflicting cases	8,691	–	10,028
	Number of unique cases	6,314	–	4,994
Duke-2	Number of conflicting cases	7,839	0	8,687
	Number of unique cases	7,511	9,489	6,295

obtained five pre-processed data sets without any missing attribute values.

To reduce error rate during classification we used a very special discretization. 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 bigger values than v in the sorted list. Our approach to discretization is the most cautious since, in the training data set, we put only one attribute value in each interval. For testing data sets, values were replaced by the corresponding intervals taken from the training data set. It could happen that a few values come into the same interval.

Surprisingly, four out of five training data sets, after replacing missing attribute values by actual attribute values and by applying our cautious discretization, were inconsistent. The training data sets are described by Table 2.

For inconsistent training data sets only possible rule sets were used for classification. Certain rules, as follows from [8], usually provide a greater error rate. Rule induction was a time-consuming process. On a DEC Alpha 21164 computer, with 512 MB of RAM, 533 MHz clock speed, rule sets were induced in elapsed real time between 21 (for Duke-2 processed by the concept closest fit algorithm) and 133 hours (for Duke-2 processed by the global concept fit algorithm). Some statistics about rule sets are presented in Table 3.

As follows from Table 3, as a result of unbalanced data sets, the average rule strength for rules describing fullterm birth is much greater than the corresponding rule strength for preterm birth. Consequently, the error rate on the original rule sets is not a good indicator of the quality of a rule set, as follows from [9].

Our basic concept is the class of preterm cases. Hence the set of all correctly predicted preterm cases are called true-positives, incorrectly predicted preterm cases (i.e., predicted as fullterm) are called false-negatives, correctly predicted fullterm cases are called true-negatives, and incorrectly predicted fullterm cases are called false-positives.

Sensitivity is the conditional probability of true-positives given actual preterm birth, i.e., the ratio of the number of true-positives to the sum of the number of true-

Table 3. Rule sets

			Global closest fit	Concept closest fit	Most common value
Duke-1	Number of rules	Preterm	734	–	618
		Fullterm	710	–	775
	Average strength of rule set	Preterm	4.87	–	8.97
		Fullterm	39.08	–	44.73
Duke-2	Number of rules	Preterm	1,202	483	1,022
		Fullterm	1,250	583	1,642
	Average strength of rule set	Preterm	2.71	9.69	4.60
		Fullterm	15.8	43.99	11.37

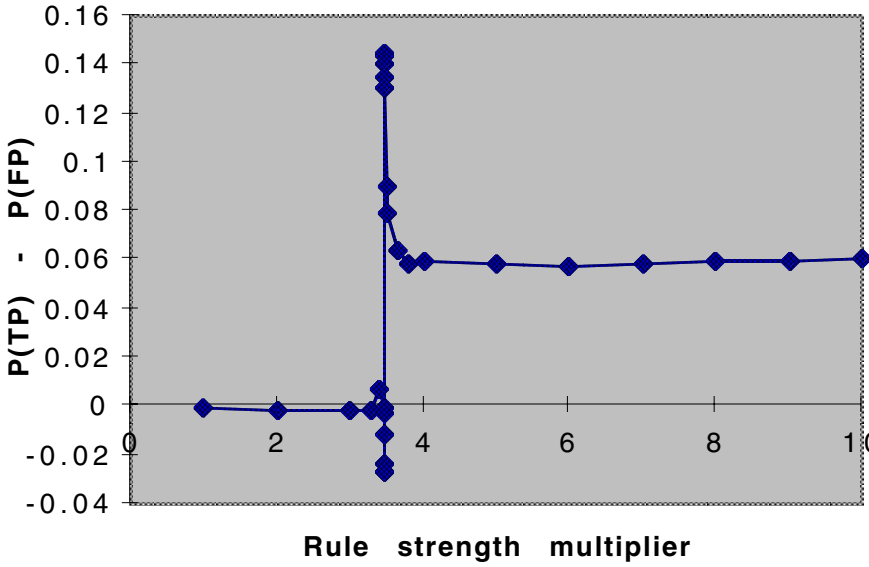

Fig. 1. P(TP) – P(FP) versus rule strength multiplier for Duke-2 data set and most common value method used for replacing missing attribute values

positives and false-negatives. It will be denoted by P(TP), following notation from [15]. Specificity is the conditional probability of true-negatives given fullterm birth, 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 birth, and equal to 1 – P(TP), will be denoted by P(FN) and the conditional probability of false-positives, given actual fullterm birth, and equal to 1 – P(TN), will be denoted by P(FP). Obviously,

$$\text{Sensitivity} + \text{Specificity} = P(TP) - P(FP) + 1,$$

so all conclusions drawn from the observations of the sum of sensitivity and specificity can be drawn from observations of P(TP) – P(FP). Another study of the sum of sensitivity and specificity was presented in [1].

Following [9], we computed the maximum of the difference between the conditional probabilities for true-positives given actual preterm birth and false-positives given actual fullterm birth as a function of the rule strength multiplier for the preterm rule set. A representative chart is presented in Fig. 1. For completeness, a typical chart (Fig. 2) shows how the true-positive, true-negative and total error rate change as a function of the rule strength multiplier. The total error rate is defined as the ratio of the number of true-positives and true-negatives to the total number of testing cases.

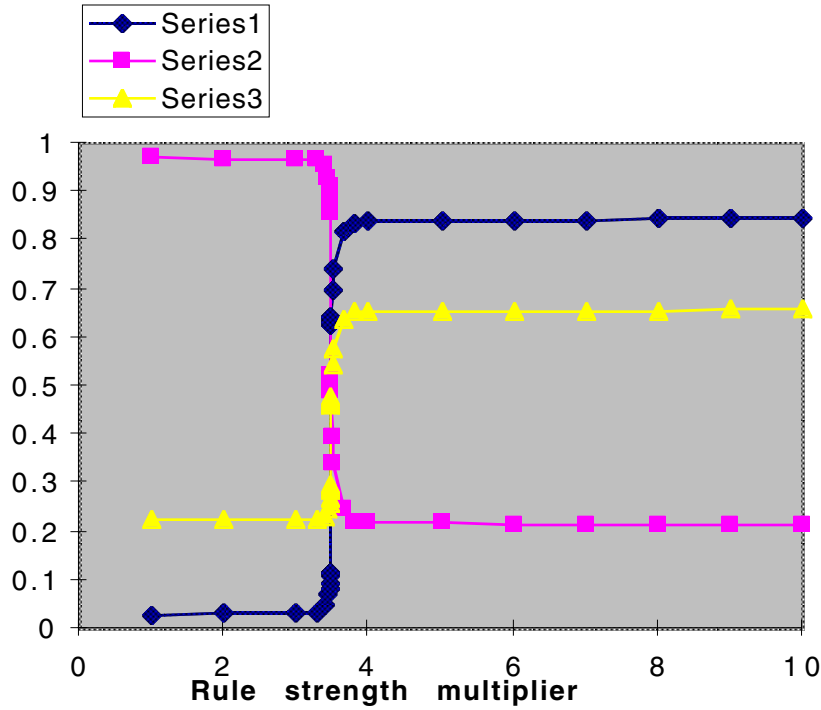

Fig. 2. Sensitivity (series 1), specificity (series 2), and total error rate (series 3) versus rule strength multiplier for Duke-2 data set and most common value method used for replacing missing attribute values

Again, following the idea from [9], in our experiments we were increasing the strength multiplier for each five rules describing preterm birth and observed P(TP) – P(FP). For each rule set, there exists some value of the rule strength multiplier, called *critical*, for which the values of P(TP) – P(FP) jumps from the minimal value to the maximal value. The respective values of true positives, true negatives, etc., and the total error rate, are also called *critical*. The results are summarized in Table 4. The total error rate, corresponding to the rule strength multiplier equal to one, is called *initial*.

The corresponding values of P(TP) – P(FP) are presented in Table 4. The critical total error rate from Table 4 is computed as the total error rate for the maximum of P(TP) – P(FP).

4 Conclusions

In our experiments the only difference between the five rule sets used for diagnosis of preterm birth is handling the missing attribute values. The maximum of the sum of sensitivity and specificity (or the maximum of P(TP) – P(FP)) is a good

Table 4. Results of experiments

	Global closest fit		Concept closest fit	Most common value	
	Duke-1	Duke-2	Duke-2	Duke-1	Duke-2
Initial total error rate	21.67	21.93	20.75	22.15	22.27
Critical total error rate	68.48	64.09	54.30	42.40	45.88
Maximum of P(TP) – P(FP)	3.65	5.97	11.69	17.07	14.43
Minimum of P(TP) – P(FP)	–15.96	–11.28	–5.37	–3.52	–2.67
Critical number of true-positives	882	838	747	615	639
Critical number of true-negatives	692	955	1535	2261	2063
Critical rule strength multiplier	8.548	6.982	6.1983	6.1855	3.478

indicator of usefulness of the rule set for diagnosis of preterm birth. It is the most important criterion of quality of the rule set. In terms of the maximum of the sum of sensitivity and specificity (or, equivalently, the maximum of P(TP) – P(FP)), the best data sets were processed by the most common value algorithm for missing attribute values. Note that the name of the algorithm is somewhat misleading because, in our experiments, we used this algorithm to compute the most common attribute value for each concept separately. The next best method is the concept closest fit algorithm. The worst results were obtained by the global closest fit.

The above ranking could be discovered not only by using the criterion of the maximum of the sum of sensitivity and specificity but also by using other criteria, for example, the minimum of the sum of sensitivity and specificity, the number of critical true-positive cases, critical false-positive cases, etc.

The initial total error rate is a poor indicator of the performance of an algorithm for handling missing attribute values. Similarly, the number of conflicting cases in the input data is a poor indicator.

Finally, it can be observed that the smaller values of the minimum of P(TP) – P(FP) correspond to the smaller values of the maximum of P(TP) – P(FP), so that the sum of the absolute values of these two numbers is roughly speaking constant.

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] Booker, L. B., Goldberg, D. E., and Holland, J. F.: Classifier systems and genetic algorithms. In Machine Learning. Paradigms and Methods. Carbonell, J. G. (ed.), The MIT Press, 1990, 235–282.

[3] Grzymala-Busse, J. W.: 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 16–19, 1991, 368–377, Lecture Notes in Artificial Intelligence, vol. 542, 1991, Springer-Verlag.

[4] 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.

[5] Grzymala-Busse, J.W. and Goodwin, L.K.: Predicting preterm birth risk using machine learning from data with missing values. Bull. of Internat. Rough Set Society **1** (1997) 17–21.

[6] Grzymala-Busse, J. W.: LERS—A knowledge discovery system. In Rough Sets in Knowledge Discovery 2, Applications, Case Studies and Software Systems, ed. by L. Polkowski and A. Skowron, Physica-Verlag, 1998, 562–565.

[7] Grzymala-Busse, J.W. and Wang A.Y.: Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values. Proc. of the Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Joint Conference on Information Sciences (JCIS'97), Research Triangle Park, NC, March 2–5, 1997, 69–72.

[8] 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.

[9] Grzymala-Busse, J. W., Goodwin, L.K., and Zhang, X.: Increasing sensitivity of preterm birth by changing rule strengths. Submitted for the 8th Workshop on Intelligent Information Systems (IIS'99), Ustronie, Poland, June 14–18, 1999.

[10] Holland, J. H., Holyoak K. J., and Nisbett, R. E.: Induction. Processes of Inference, Learning, and Discovery. The MIT Press, 1986.

[11] Michalski, R. S., Mozetic, I., Hong, J. and Lavrac, N.: The AQ15 inductive learning system: An overview and experiments. Department of Computer Science, University of Illinois, Rep. UIUCDCD-R-86-1260, 1986.

[12] Pawlak, Z.: Rough sets. International Journal Computer and Information Sciences **11** (1982) 341–356.

[13] Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, 1991.

[14] Stefanowski, J.: On rough set based approaches to induction of decision rules. In Polkowski L., Skowron A. (eds.) Rough Sets in Data Mining and Knowledge Discovery, Physica-Verlag, 1998, 500–529.

[15] Swets, J.A. and Pickett, R.M.: Evaluation of Diagnostic Systems. Methods from Signal Detection Theory. Academic Press, 1982.

[16] Ziarko, W.: Systems: DataQuest, DataLogic and KDDR. Proc. of the Fourth Int. Workshop on Rough Sets, Fuzzy Sets and Machine Discovery RSFD'96, Tokyo, Japan, November 6–8, 1996, 441–442.