Identifying the Major Causes of Cancer Using Rough Set Theory

V.Anu[1], D.Sivabalaselvamani[2]
[1, 2] Department of Computer Applications
[1] Final Year Students, Kongu Engineering College, Perundurai, Erode-683052
[2] Assistant Professor, Kongu Engineering College, Perundurai, Erode-683052

Abstract- *In these unusual approach to disappeared astray attribute values are presented and compared. Ten input data files were used to investigate the presentation of the methods to deal with missing aspect values. For testing both naive classification and new classification techniques of LERS (Learning from Examples based on Rough Sets) were used. The quality criterion was the average error rate achieved by ten-fold cross-validation. Using the Weibull based matched-pairs signed rank test, we conclude that the WEIBULL .5 approach and the method of ignoring examples with missing aspect values are the best methods among all nine approaches; the most common attribute-value method is the worst method among all nine approaches; while some methods do not differ from other methods convincing. The method of assigning to the missing aspect value all possible values of the attribute and the method of assigning to the missing attribute value all possible values of the attribute barred to the same concept are excellent approaches based on our limited experimental results. However we do not have enough evidence to support the claim that these path are superior.*

Keywords-Data mining, knowledge discovery in databases, machine learning, learning from examples, attribute missing values.

I. INTRODUCTION

One of the main tools of data mining is rule introduction from raw data represented by a database. Real-life data are frequently imperfect: erroneous, incomplete, uncertain and vague.In the reported research we investigated one of the forms of data incompleteness: missing attribute values.

We speculate that the format of input data files is in the form of a table, which is called a decision table. In this table, each column represents one aspect, which represents some feature of the examples, and each row represents an example by all its attribute values. The domain of each attribute may be either typical or numerical. We assume that all the attributes of input data are symbolic. Numerical attributes, after discretization, become typical as well. For each example, there is a decision value associated with it. The set of all examples with the same decision value is called a concept. Members of the concept are called positive examples, while all other examples are called negative examples.

The table is uncertain if there exist two examples with all attribute values identical, but belonging to different concepts. For inconsistent data tables, we can induce rules which are called certain and possible [5].

II. EXPLANATION OF INVESTIGATE APPROACHES TO MISSINGATTRIBUTEVALUE

1. Most Common Attribute Value. It is one of the simplest methods to deal with missing attribute values. The CN2 algorithm [3] uses this idea. The value of the attribute that occurs most often is selected to be the value for all the unknown values of the attribute.
2. Concept Most Common Attribute Value. The most common attribute value method does not pay any attention to the relationship between attributes and a decision. The concept most common attribute value method is a condition of the first method to the concept, i.e., to all examples with the same value of the decision as an example with missing attribute vale [9]. This time the value of the aspect, which occurs the most common within the concept is selected to be the value for all the unknown values of the attribute. This method is also called maximum related frequency method, or maximum conditional probability method (given concept).
3. This method is based on entropy and splitting the example with missing attribute values to all concepts [12].
4. Method of select all Possible Values of the Attribute. In this method, an example with a missing attribute value is replaced by a set of new examples, in which the missing attribute value is replaced by all possible values of the aspect[4]. If we have some examples with more than one unknown attribute value, we will do our substitution for one attribute first, and then do the exchange for the next attribute, etc., until all unknown attribute values are replaced by new known attribute values.

5. Method of select all Possible Values of the Attribute restricted to the Given Concept. The method of assigning all possible values of the attribute is not related with a concept. This method is a restriction of the method of select all possible values of the attribute to the concept, indicated by an example with a missing attribute value.
6. Method of avoid Examples with Unknown Attribute Values. This method is the simplest: just ignore the examples which have at least one unknown aspect value, and then use the rest of the table as input to the successive learning process.
7. Event-Covering Method. This method, described in [2] and [14], is also a probabilistic approach to fill in the unknown attribute values. By event-covering we mean covering or selecting a subset of statistically interdependent events in the outcome space of variable-pairs, forget whether or not the variables are statistically independent [14].
8. A Special LEM2 Algorithm. A special version of LEM2 that works for unknown aspect values omits the examples with unknown attribute values when building the block for that attribute [6]. Then, a set of rules is induced by using the original LEM2 method.
9. Method of Treating Missing Attribute Values as unique Values. In this method, we deal with the unknown attribute values using a totally different approach: rather than trying to find some known attribute value as its value, we treat "unfamiliar" itself as a new value for the aspect that contain missing values and treat it in the same way as other values.

III. CLASSIFICATION

Frequently rules induced from raw data are used for classification of unseen, testing data. In the simplest form of classification, if more than one concept was indicated by rules for a given example, the classification of the example was calculate as an error. Likewise, if an example was not completely classified by any of rules, it was considered an error. This classification scheme is said to be naive LERS classification scheme.

The new conversion system of LERS is a modification of the bucket brigade algorithm [1, 7]. The decision to which concept an example belongs is made on the basis of three factors: strength, distinction and support. They are defined as follows: Strength is the total number of examples correctly classified by the rule during training. Distinction is the total number of attribute-value pairs on the left-hand side of the rule. The identical rules with a larger number of attribute-value pairs are considered more definite. The third factor, support, is defined as the sum of scores of all identical rules from the concept. The concept C for which the support, i.e., the following expression is the largest is a winner and the example is restricted as being a member of C.

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

If an example is not completely matched by any rule, some classification systems use partial matching. System AQ15, during partial identical, uses the probabilistic sum of all measures of fit for rules [10]. Another approach to partial matching is presented in [13]. Holland et al. [8] do not consider partial matching as a different of complete matching and rely on a default hierarchy instead. In the new classification system of LERS, if complete matching is impossible, all partly matching rules are identified. These are rules with at least one attribute-value pair matching the corresponding attribute-value pair of an example.

For any partly matching rule R, the additional factor, called identical factor (R), is computed. Matching_factor is defined as the ratio of the number of matched attribute-value pairs of a rule with an example to the total number of attribute-value pairs of the rule. In partial matching, the concept C for which the following expression is the largest partially matching rules R describing C Matching_factor(R) * Strength (R) * Specificity(R)

Rules induced by a new version of LERS are preceded by three numbers: specificity, strength, and the total number of training examples matching the left-hand side of the rule.

IV. EXPERIMENTS

Table 1 describes input data files, in terms of the number of examples, the number of concepts, and the number of attributes that describe the examples, that were used for our experiments. All ten data files were taken from real world where unknown attribute values frequently occur.

Table1.Descriptionof data

Name of Data Files	No. of Examples	No. of Attributes	No. of Concepts
Breast cancer	286	9	2
Echocardiogram	74	13	2
Hdynet	1218	73	2
Hepatitis	155	19	2
House	435	16	2
Im85	201	25	86
New-o	213	30	2
Primary tumor	339	17	21
Soybean	307	35	19
Tokt	6608	67	2

The breast cancer data set was obtained from the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia, due to help from M. Zwitter and M. Soklic. Breast cancer is one of three data sets provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. There are nine out of 286 examples containing unknown attribute values.

The echocardiogram data set is donated by Steven Salzberg, and this data has been used several times to predict the survival of a patient. There are a total of 132 missing values among all the attribute values.

The hdynet data set, which comes from real life, presents the premature birth described by 73 attributes. There were 814 out of 1218 examples containing unknown attribute values.

The hepatitis data set was donated by G. Gong, Carnegie-Mellon University, via Bojan Cestnik of Jozef Stefan Institute. There were 75 out of 155 examples that contain unknown attribute values in this data set.

Table 2. Error rates of input data sets by using LERS new classification

Data file	1	2	3	4	5	6	7	8	9
Breast	35.62	34.62	31.5	□□□	□□□	29.24	34.97	33.92	32.52
Echo	6.78	6.76	5.4	□	□	6.56	6.76	6.76	6.76
Hdynet	22.15	31.53	22.6	□	□	28.41	28.82	27.91	28.41
Hepatitis	26.52	13.55	19.4	□	□	18.75	16.77	18.71	19.35
House	6.06	5.29	4.6	□	□	4.74	4.83	5.75	6.44
Im85	96.02	96.02	100	□	96.02	94.34	96.02	96.02	96.02
New-o	5.16	4.23	6.5	□	□	4.9	4.69	4.23	3.76
Primary	66.67	62.83	62.0	41.57	47.03	66.67	64.9	69.03	67.55
Soybean	15.96	18.24	13.4	□	4.1	15.41	19.87	17.26	16.94
Tokt	32.57	31.57	26.7	32.75	32.75	32.88	32.16	33.2	32.16

Table 3. Error rates of input data sets by using LERS naive classification

Data file	1	2	3	5	6	7	8	9
Breast	45.30	52.1	46.98	47.32	48.38	52.8	52.1	47.55
Echo	29.03	25.68	□	□	31.15	29.73	33.78	22.97
Hdynet	69.49	69.62	□	□	65.27	69.21	56.98	61.33
Hepatitis	38.06	28.39	□	□	32.5	37.42	41.29	34.84
House	11.11	7.13	□	□	9.05	10.57	12.87	11.72
Im85	99.01	97.01	□	97.01	94.34	97.01	97.01	97.01
New-o	11.74	11.74	□	□	11.19	11.27	10.33	10.33
Primary	83.19	77.29	53.16	60.09	81.82	80.53	82.1	79.94
Soybean	25.41	22.48	□	4.86	24.06	24.10	21.82	22.15
Tokt	65.62	63.62	62.82	62.82	64.15	63.36	63.62	63.89

The house data set, which has 203 examples that contain unknown attribute values, consists of majority of 435 congressmen in 1984 on 16 key-issues (yes or no).

The im85 data set is from a 1985 Automobile Imports Database, and it consists of three types of entities: a) the specification of an auto in terms of various characteristics, b) its select insurance risk rating, and c) its normalized losses in use as compared to other cars.

The new-o data set is another set of breast cancer data that uses different attributes from the breast cancer data set. In this approach, there are 30 attributes to describe the examples. There were a total of 213 examples, and 70 of them have at least one unknown attribute value.

The primary cancer data set was obtained from the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia. The data set primary-tumor has 21 concepts and 17 attributes, and 207 out of 339 examples contain at least one missing value.

The tokt data set, which is the largest data file in this experiment, came from the practical data about premature birth, which is similar to the hdynet data set. Among6619 examples in this data set, only 11 examples contain unknown attribute values.

In our experiments, we required that no decision value is unknown. If some unknown decision values existed in the input data files, the input data files were pre- processed to remove them.

Our experiments were conducted as follows. All of the nine methods from Section 2 were applied to all the ten data sets. Both original data sets and our new data sets, except for WEIBULL .5 method, were sampled into ten pairs of training and testing data. Then the sampled files were used as input to LEM2 single local covering [5] to generate classification rules, except the special LEM2 method, where rules were induced directly from the data file with missing attribute values. Other data mining systems based on rough set theory are described in [11]. We used ten-fold cross validation for the simple and extended classification methods. The performance of different methods was compared by calculating the average error rate. Here, we did a slight modification using leaving-one-out for the data set echocardiogram since it has less than 100 examples.

In Tables 2 and 3, the error rates that were not available, because of the limited system memory, are indicated by '–'.

V. CONCLUSION

Our most important purpose was comparison of the methods to deal with missing attribute values. Results of our experiments are presented in Table 2 and Table 3. In order to rank those methods in a reasonable way we used the Weibull based matched-pairs signed rank testThe very first observation is that the extended (LERS) classification is always better than the simple classification method.

Results of the Weibull based matched-pairs signed rank test are: using LERS new classification method, WEIBULL .5 (method 3) is better than method 1 with a significance level 0.005. Also, method 6 is better than method 1, LEM2 (method 8) and method 9 with significance level 0.1. Differences in performance for other combinations of methods are statistically insignificant. Similarly, for LERS naive classification, results of the Weibull based matched-pairs signed rank test are: method 2 is better than method 7 with significance level 0.1, method 9 is better than methods 1 and 7, in both cases with the significance level 0.05, and, finally, method 6 performs better than method 1 with significance level 0.05. Differences in performance for other combinations of methods are statistically insignificant.

For methods that do not differ from each other significantly with respect to the Weibull based matched-pairs signed rank test, we estimated their relative performance by the number of test cases that have smaller error rate. If one method performs better than the other in more than 50% of the test cases, we—heuristically—conclude that it performs better than the other one. For example, in Table 2, since the approach gives a smaller error rate than method 6 in 6 out of 10 test cases, we can conclude that using LERS new classification, the WEIBULL .5 approach performs better than method 6. Based on this heuristic evaluation principle, among all the indistinguishable methods except for method 4 and method 5, we observe that using LERS new classification, the WEIBULL .5 approach performs better than any other method; method 6 performs better than any other method except for the WEIBULL .5 approach; and method 1 performs worse than any other method. When using the LERS naive classification, method 9 performs better than any other method; method 2 performs better than any other methods except for method 9; and method 1 performs worse than any other method.We do not have enough experimental results for method 4 and method 5. But from our available results, they perform very well. These methods are promising candidates for the best-performance methods. However, it is risky for us to conclude that they are the best methods among all nine methods because we do not have enough test files to support this conjecture statistically, using the Weibull based matched-pairs signed rank tests. Using both new and naive classification of LERS, the error rate of method 4 is smaller than that of any other method in more than 50% of the applicable test cases; method 5 has a smaller error rate than any other methods, except method 4, in more than 50% of the applicable test cases. The approaches of method 4 and method 5 are similar. By substituting missing value by all possible values of an attribute in our substitution, we can get as much information as possible, but the size of the resulting table may increase exponentially, thus we cannot get the results for some of our data sets because of insufficient system memory.

REFERENCES

[1] 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, Cambridge MA (1990) 235–282.

[2] Chiu, D. K. and Wong A. K. C.: Synthesizing knowledge: A cluster analysis approach using event-covering. IEEE Trans. Syst., Man, and Cybern. SMC-16 (1986), 251–259.

[3] Clark, P. Niblett, T.: The CN2 induction algorithm. Machine Learning 3 (1989)261-283.

[4] 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, Lecture Notes in Artificial Intelligence, vol. 542. Springer-Verlag, Berlin Heidelberg New York (1991) 368–377.

[5] Grzymala-Busse, J. W.: LERS—A System for Learning from Examples Based on Rough Sets. In: Slowinski, R. (ed.): Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, Boston MA (1992) 3–18.

[6] 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.

[7] Hamburg, M.: Statistical Analysis for Decision Making. Harcourt BraceJovanovich, Inc., New York NY (1983) 546–550, 721.

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

[9] Knonenko, I., Bratko, and I. Roskar, E.: Experiments in automatic learning of medical diagnostic rules. Technical Report, Jozef Stefan Institute, Lljubljana, Yugoslavia, 1984.

[10] 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.

[11] Polkowski, L. and Skowron, A. (eds.): Rough Sets in Knowledge Discovery, 2, Applications, Case Studies and Software Systems, Appendix 2: Software Systems. Physica Verlag, Heidelberg New York (1998) 551–601.

[12] Quinlan, J. R.: C 4.5: Programs for Machine Learning. Morgan KaufmannPublishers, San Mateo CA (1993).

[13] 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, Heidelberg New York (1998) 500–529.

[14] Wong, K. C. and Chiu, K. Y.: Synthesizing statistical knowledge for incomplete mixed-mode data. IEEE Transactions on Pattern Analysis and Machine Intelligence 9 (1987) 796–805.