Comparisons on Different Approaches to Assign Missing Attribute Values

Jiye Li[1] and Nick Cercone[2]

[1] School of Computer Science, University of Waterloo
200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1
j27li@uwaterloo.ca
[2] Faculty of Computer Science, Dalhousie University
6050 University Avenue, Halifax, Nova Scotia, Canada B3H 1W5
nick@cs.dal.ca

Abstract. A commonly-used and naive solution to process data with missing attribute values is to ignore the instances which contain missing attribute values. This method may neglect important information within the data, significant amount of data could be easily discarded, and the discovered knowledge may not contain significant rules. Some methods, such as assigning the most common values or assigning an average value to the missing attribute, may make good use of all the available data. However the assigned value may not come from the information which the data originally derived, thus noise is brought to the data. We introduce a new approach **RSFit** on processing data with missing attribute values based on rough sets theory. By matching attribute-value pairs among the same core or reduct of the original data set, the assigned value preserves the characteristics of the original data set. We compare our approach with "closest fit approach globally" and "closest fit approach in the same concept". Experimental results on UCI data sets and a real geriatric care data set show our approach achieves comparable accuracy on assigning the missing values while significantly reduces the computation time.

Keywords: Missing Attribute Values, Rough Sets Theory, Core, Reduct

1 Introduction

Missing attribute values are commonly existing in real-world data set. They may come from the data collecting process, or redundant diagnose tests, unknown data and so on. Discarding all data containing the missing attribute values cannot fully preserve the characteristics of the original data. If we understand the background knowledge or the data collection process, and use the original context to assign the missing values, we will have the best approach for handling missing attribute values. But in reality, it is difficult to know the original meaning for this missing data. Various approaches on how to cope with the missing attribute values have been proposed in the past years. In [1] nine approaches on filling in the missing attribute values were introduced, such as selecting the "most common attribute value", the "concept most common attribute value",

"assigning all possible values of the attribute restricted to the given concept", "ignoring examples with unknown attribute values", "treating missing attribute values as special values", "event-covering method" and so on. Experiments on ten data sets were conducted to compare the performances. In [2] a "closest fit" approach was proposed to compare the vectors of all the attribute pairs from a preterm birth data set, and assign the value from the most similar pair to the missing value. In a more recent effort [3] four interpretations on the meanings of missing attribute values such as "lost" values and "do not care" values are discussed. Different approaches from rough sets theory are demonstrated on selecting values for the individual interpreted meanings.

We investigate the effectiveness of assigning missing attribute values from rough sets perspective. Rough sets theory, proposed in the 1980's by Pawlak [4], has been used for attribute selection, rule discovery and many knowledge discovery applications in the areas such as data mining, machine learning and medical modeling. Core and reduct are among the most important concepts in this theory. A reduct contains a subset of condition attributes that are sufficient enough to represent the whole data set. The intersection of all the reduct is the core. Therefore by examining only attributes within the same core or reduct for the matched or similar attribute-value pairs for the data instance containing the missing attribute values, we assign the most relevant value for the missing attribute. This is because attributes in the same core or reduct are much more related to each other than attributes from all the data set, and they keep the integrity of the original data set. The assigned missing values therefore come within the data itself. Less noise will be brought into the data set, and the computational cost is also less than that of the "closest fit" approach. Experiments on UCI data sets and a geriatric care data set demonstrate our proposed approach on assigning missing attribute values can greatly reduce the computation time and at the same time maintain a satisfactory accuracy.

The rest of the report is organized as follows. Our proposed rough sets based approach is explained in section 2. Experiment design and the comparison results are described in section 3. Section 4 gives conclusion remarks and discuss future work.

2 RSFit Approach to Assign Missing Values

We first make definitions to be used in the following description of our approach. The input to our approach is a decision table $T = (C, D)$, where $C = \{c_1, c_2, \ldots, c_m\}$ is the condition attribute set, and $D = \{d_1, d_2, \ldots, d_l\}$ is the decision attribute set. $U = \{u_1, u_2, \ldots, u_n\}$ represent the set of data instances in T. For each u_i $(1 \leq i \leq n)$, an **attribute-value pair** for this data instance is defined to be $u_i = (v_{1i}, v_{2i}, \ldots, v_{mi}, d_i)$, where v_{1i} is the attribute value for condition attribute c_1, v_{2i} is the attribute value for condition attribute c_2, ..., v_{mi} is the attribute value for condition attribute c_m.

2.1 Detailed Explanation

The core or the reduct of a data set contains a set of attributes that are able to represent the original data set. The attributes contained in the same core or the reduct set are dependent on each other to a certain statistic measure. We consider attribute-value pairs contained in the same core or reduct set to find the best match for the missing values. This approach is inspired by the "closest fit" approach by Grzymala-Busse [2], however it is different from it. Instead of searching the whole data set for closest matched attribute-value pairs, RSFit searches only the attribute-value pairs within the core or the reduct. The attribute-value pair is defined as following.

For each missing attribute value, we let the attribute be the "target attribute"(represented as c_k in the following). We assume that missing attribute values are only existing in the condition attributes not in the decision attributes. We explain our approach in detail on how to find the matched value for this target attribute.

Firstly, we obtain the core of the data set $T = (C, D)$ based on the core algorithm introduced in [5]. If the target attribute c_k does not belong to the core, we include c_k into the core. In case there is no core for T, we consider the reduct of T. ROSETTA software [6] is used for reduct generation. There are a few reduct generation algorithms provided by ROSETTA. We use Johnson's algorithm for single reduct generation. In case of no reducts containing the target attribute c_k, we include the target c_k into the reduct.

Secondly, a new decision table $T' = (C', D)$ is created based on the previous step, where $C' = \{c'_1, c'_2, \ldots, c_k, \ldots, c'_{m'}\}$, $1 \leq m' \leq m$, $1 \leq k \leq m'$, and $C' \subseteq C$, C' is either the core or the reduct of C, $U' = \{u_1, u_2, \ldots, u_{n'}\}$, $1 \leq n' \leq n$. There are two possibilities for selecting the data instances. One possibility is to include other data instances with missing values to predict the current target attribute value; the other option is to exclude all the other data instances containing missing attribute values. We allow the other missing attribute values existing by designing the proper match function.

Thirdly, in T', when considering the match cases, there are two possibilities existing. One possibility is that we consider all the data instances; the other is to consider data instances having the same decision attribute values while finding a matched attribute-value pair. Here we call the first possibility *global*, and the second *concept*. We perform experiments to test both possibilities in our paper. Fourthly, we define the distance function to compute the similarities between different attribute-value pairs. The details of the distance function is elaborated in the following. Let $u_i = (v_{1i}, v_{2i}, \ldots, v_{ki}, \ldots, v_{m'i}, d_i)$ $(1 \leq i \leq n')$ be the attribute-value pair containing the missing attribute value v_{ki} (represented as $v_{ki} =?$) for c_k $(1 \leq k \leq m')$. Distance functions, such as Euclidean distance and Manhattan distance, are used in the instance-based learning to compare the similarity between a test instance and the training instances [7]. We use

Manhattan distance [3]to evaluate the distance between an attribute-value pair containing missing attribute values with other attribute-value pairs. This formula is also used in the "closest fit" approach [2]. Let u_j be a data instance from U. The distance between u_j to the target data instance u_i is defined as[4]

$$distance(u_i, u_j) = \frac{|v_{i1} - v_{j1}|}{\max v_1 - \min v_1} + \frac{|v_{i2} - v_{j2}|}{\max v_2 - \min v_2} + \ldots + \frac{|v_{im} - v_{jm}|}{\max v_m - \min v_m}.$$

For attributes which are the missing attribute values, the distance is set to be 1, which specifies the maximum difference between unknown values. The best match has the smallest difference from the target attribute-value pair. After the best matched attribute-value pair is returned by the algorithm, the corresponding value will be assigned to the target attribute. We consider all the attributes as numerical attributes. In case of symbolic attributes, we convert them to numerical ones during the preprocessing stage.

In case there are multiple matched attribute-value pairs for the missing attribute, one of the values is randomly selected to be assigned to the missing value.

2.2 Evaluation Method

Our goal is to test the accuracy of using our method to predict the missing values, and compare the accuracy and the computation time with "closest fit global" and "closest fit concept" approaches. We use the following way to perform the evaluation process. We consider complete data sets as our input. For each data set, we randomly select a certain number of missing values among the whole data set to produce n missing attribute values per data set. We test different approaches on assigning the missing values, and compare the accuracy of the prediction. In order to average the odd of the randomly selected missing attributes, we perform this process 100 times for each data set and average the accuracy.

3 Experiment

3.1 A Walk Through Example

We demonstrate our approach by an artificial car data set which appeared in [5] shown in Table 1. One missing attribute value is randomly selected across the data set as shown by Table 2.

Firstly, the core is obtained for this data set as "Make_model" and "trans". Since the core attributes exist, the missing attribute "compress" does not belong to the core, we add "compress" to the core set. The new data set containing only the core, target attribute "compress" and the decision attribute are shown in Table 3. Then we find the match for attribute "compress" in u_8. For "RSFit

[3] In our experiments, the prediction results by Manhattan distance and Euclidean distance returned the same accuracy. Because the computation for Manhattan distance is faster, we use Manhattan distance as the distance function.

[4] In the algorithm, $|x|$ returns the absolute value of x.

Table 1. Artificial Car Data Set

U	Make_model	cyl	door	displace	compress	power	trans	weight	mileage
1	usa	6	2	medium	high	high	auto	medium	medium
2	usa	6	4	medium	medium	medium	manual	medium	medium
3	usa	4	2	small	high	medium	auto	medium	medium
4	usa	4	2	medium	medium	medium	manual	medium	medium
5	usa	4	2	medium	medium	high	manual	medium	medium
6	usa	6	4	medium	medium	high	auto	medium	medium
7	usa	4	2	medium	medium	high	auto	medium	medium
8	usa	4	2	medium	high	high	manual	light	high
9	japan	4	2	small	high	low	manual	light	high
10	japan	4	2	medium	medium	medium	manual	medium	high
11	japan	4	2	small	high	high	manual	medium	high
12	japan	4	2	small	medium	low	manual	medium	high
13	japan	4	2	small	high	medium	manual	medium	high
14	usa	4	2	small	high	medium	manual	medium	high

Table 2. Artificial Car Data Set with One Missing Attribute Values

U	Make_model	cyl	door	displace	compress	power	trans	weight	mileage
1	usa	6	2	medium	high	high	auto	medium	medium
2	usa	6	4	medium	medium	medium	manual	medium	medium
3	usa	4	2	small	high	medium	auto	medium	medium
4	usa	4	2	medium	medium	medium	manual	medium	medium
5	usa	4	2	medium	medium	high	manual	medium	medium
6	usa	6	4	medium	medium	high	auto	medium	medium
7	usa	4	2	medium	medium	high	auto	medium	medium
8	usa	4	2	medium	?	high	manual	light	high
9	japan	4	2	small	high	low	manual	light	high
10	japan	4	2	medium	medium	medium	manual	medium	high
11	japan	4	2	small	high	high	manual	medium	high
12	japan	4	2	small	medium	low	manual	medium	high
13	japan	4	2	small	high	medium	manual	medium	high
14	usa	4	2	small	high	medium	manual	medium	high

global", we find the u_{14} has the smallest difference from u_8, therefore u_{14} is the best match. We assign $c_{compress_{14}}$ to $c_{compress_8}$, which is "High" (correct prediction). For "RSFit concept", we only look for attribute-value pairs that have the same decision attribute value as u_8, which is $mileage = high$. We find u_{14} is the best match. We assign $c_{compress_{14}}$ to $c_{compress_8}$, which is "High" (Correct prediction). For "closest fit global" approach, we examine all the instances in the data set. u_5 is the closest fit, $c_{compress_5}$ = "Medium" (wrong prediction). For "closest fit concept" approach, we examine only the data with decision at-

Table 3. New Decision Table for Car Data Set Based on Core Set

U	Make_model	compress	trans	mileage
1	usa	high	auto	medium
2	usa	medium	manual	medium
3	usa	high	auto	medium
4	usa	medium	manual	medium
5	usa	medium	manual	medium
6	usa	medium	auto	medium
7	usa	medium	auto	medium
8	usa	?	manual	high
9	japan	high	manual	high
10	japan	medium	manual	high
11	japan	high	manual	high
12	japan	medium	manual	high
13	japan	high	manual	high
14	usa	high	manual	high

tribute "High". We find u_{10} with $c_{compress_{10}}$ = "Medium" as the match (wrong prediction).

3.2 Experiment on UCI Data Sets and a Geriatric Care Data Set

In order to test our proposed approach, we experiment on selected UCI data sets [8] and a geriatric care data set [9], which contain no missing attribute values.

These data sets can be divided into two categories. One category of data sets contain core attributes, such as, geriatric care data, spambase data and zoo data. The other set of data sets do not contain core attributes, such as lymphography data. For the type of data set in which the core attributes are all the condition attributes, we do not discuss in this paper (in this case our method is the same as the closest fit approach).

Geriatric Care Data Set We perform experiments on a geriatric care data set from Dalhousie University Department of Medical. This data set contains 8547 patient records with 44 symptoms and their survival status. The data set is used to determine the survival status of a patient given all the symptoms he or she shows. We use *survival status* as the decision attribute, and the 44 symptoms of a patient as condition attributes, which includes *education level, the eyesight, the age of the patient at investigation* and so on. [5] There is no missing value in this data set. There are 12 inconsistent data entries in the medical data set. After removing these instances, the data contains 8535 records. [6] Table 4 gives

[5] Refer to [9] for details about this data set.

[6] Notice from our previous experiments that core generation algorithm can not return correct core attributes when the data set contains inconsistent data entries.

selected data records of this data set. There are 14 core attributes generated for

Table 4. Geriatric Care Data Set

edulevel	eyesight	hearing	health	trouble	livealone	cough	hbp	heart	stroke	...	sex	livedead
0.6364	0.25	0.50	0.25	0.00	0.00	0.00	0.00	0.00	0.00	...	1	0
0.7273	0.50	0.25	0.25	0.50	0.00	0.00	0.00	0.00	0.00	...	2	0
0.9091	0.25	0.50	0.00	0.00	0.00	0.00	1.00	1.00	0.00	...	1	0
0.5455	0.25	0.25	0.50	0.00	1.00	1.00	0.00	0.00	0.00	...	2	0
0.4545	0.25	0.25	0.25	0.00	1.00	0.00	1.00	0.00	0.00	...	2	0
0.2727	0.00	0.00	0.25	0.50	1.00	0.00	1.00	0.00	0.00	...	2	0
0.0000	0.25	0.25	0.25	0.00	0.00	0.00	0.00	1.00	0.00	...	1	0
0.8182	0.00	0.50	0.00	0.00	0.00	0.00	1.00	0.00	0.00	...	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...

this data set. They are *eartroub, livealone, heart, hbp, eyetroub, hearing, sex, health, edulevel, chest, housewk, diabetes, dental, studyage.*

Lymphography Data The data set contains 148 instances and 18 condition attributes. There are no missing attribute values in this data. We check that there is no inconsistent data. The core is empty for this data set. Johnson's reduct generated from this data set contains *blockofaffere, changesinnode, changesinstru, specialforms, dislocationof, noofnodesin.*

Spambase Data This data set originally contains $4,601$ instances and 57 condition attributes. It is used to classify spam and non-spam emails. Most of the attributes indicate whether a certain word (such as, order, report) or character (such as !, #) appears frequently in the emails. There are no missing attribute values. There are 6 inconsistent data instances that are removed. The core attributes, which are essential to determine whether an email is not a spam email, are, the word frequency of "george", "meeting", 're", "you", "edu", "!", and the total number of capital letters in the email. In addition, it is interesting to pay attention to the reducts as well. They are important information on identifying the possible spam emails.

Zoo Data This artificial data set contains 7 classes of animals, 17 condition attributes, 101 data instances, and there are no missing attribute values in this data set. Since the first condition attribute "animal name" is unique for each instance, and we consider each instance a unique itemset, we do not consider this attribute in our experiment. There are no inconsistent data in this data set. The core attributes are *aquatic, legs.*

3.3 Comparison Results

The compared approaches are implemented by Perl and the experiments are conducted on Sun Fire V880, four 900Mhz UltraSPARC III processors. Our proposed rough sets based approach considers a subset of the attributes (the reduct or the core). In order to compare whether the reduct or the core provide a better choice of attributes, we also compare our approach against a random subset of the attributes. Given a reduct of size n, we randomly choose a combination of n attributes from a uniform distribution. The comparison results on processing missing attribute values between RSFit approach, closest fit approach and random approach on geriatric care data set spambase data set, lymphography data set and zoo data set are shown in Fig. 1, Fig. 2, Fig. 3, and Fig. 4. The reduct and core generation time are not included in the comparison results.

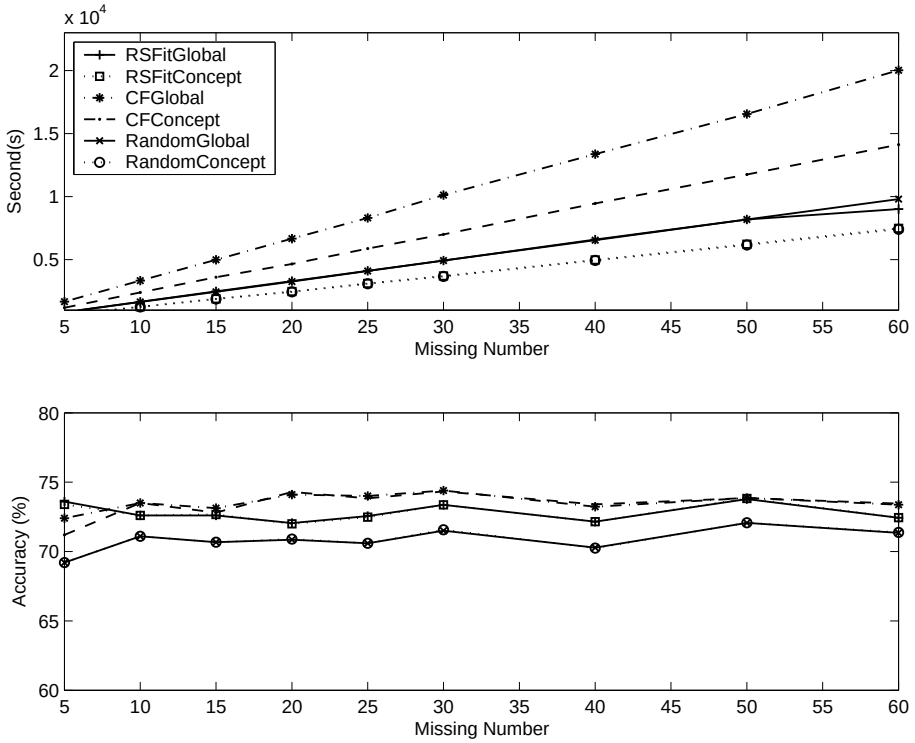

Fig. 1. Comparison Figure for Geriatric Care Data

The comparison results are shown in the following Table 5, Table 6, Table 7 and Table 8.

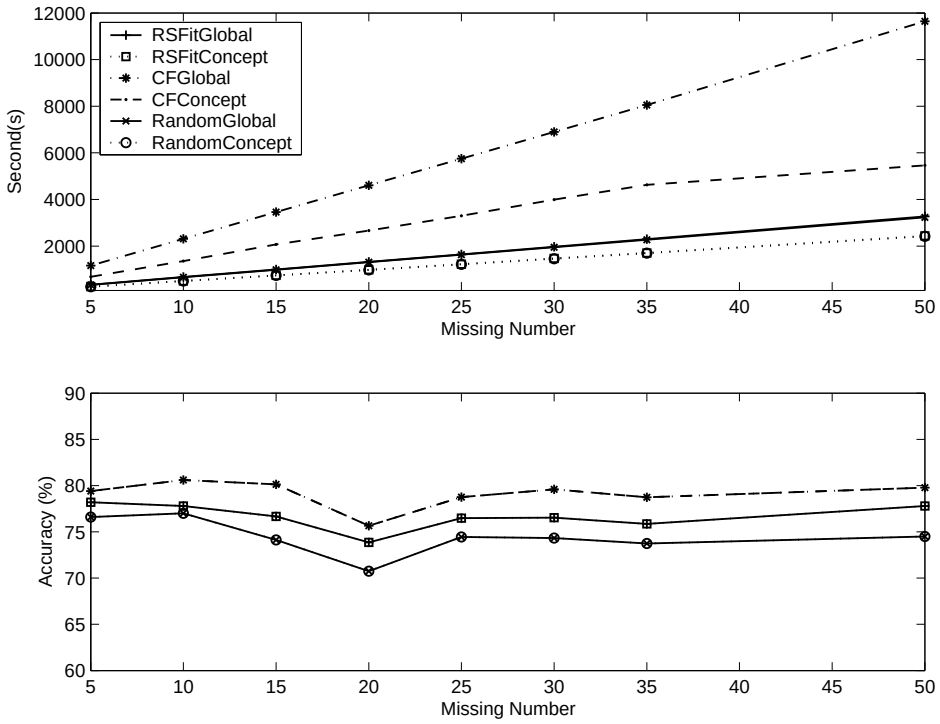

Fig. 2. Comparison Figure for Spambase Data

3.4 Discussions

In the comparison figures, "RSFitGlobal" and "RSFitConcept" stand for the new approach proposed in this paper. "CFGlobal" and "CFConcept" stand for the "closest fit" approach from [2]. "RandomGlobal" and "RandomConcept" stand for the random selected attributes approach. For each figure, the upper chart shows the prediction time, the lower chart shows the prediction accuracy. Our proposed rough sets theory based method achieved significant saving on computation time for assigning missing attribute values. It can be used in the situation when time is the most important issue, with the sacrifice of less precision. The time saving is quite noticeable for larger data sets such as geriatric care and spambase data set. Take the geriatric care data as an example, among the 44 condition attributes, we only consider 14 of them which are core attributes. Comparing "RSFitGlobal" to "CFGlobal", the prediction precision of ours is on average 0.762% lower than the "closest fit" approach, however the computation time of ours is on average 49.026% of the computation time for the "closest fit" approach. The "RSFitConcept" and "RSFitGlobal" achieve similar prediction accuracy, however, the "RSFitConcept" takes slightly less computation time because the amount of data the approach processes is less. This observation also

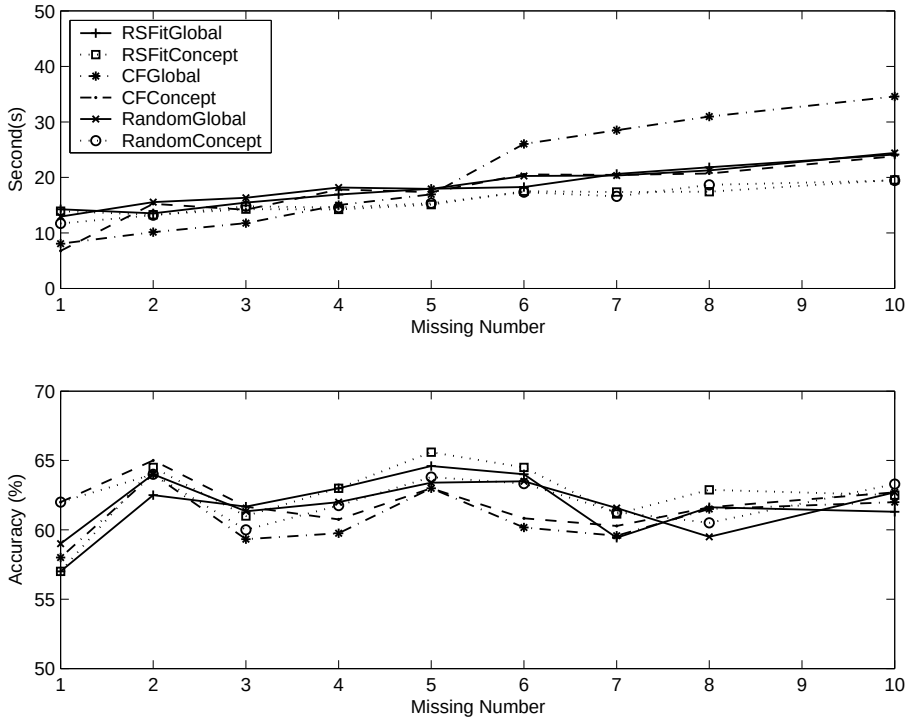

Fig. 3. Comparison Figure for Lymphography Data

applies to "closest fit" approach and random approach. The experimental results also shows that RSFit approach provides a higher prediction accuracy than the random approach. The reduct from the rough sets theory presents a better choice of attributes than the randomly selected attributes.

4 Concluding Remarks and Future Work

In this paper, we introduce a new approach to assign missing attribute values based on rough sets theory. Comparing to the "closest fit" approach proposed by Grzymala-Busse, RSFit approach significantly reduces the computation time and comparable accuracy is achieved. As future work, we are interested in using semantic approach to assign the missing values. Semantic approaches of null value problem in natural language processing database interfaces was first studied by Kao, Cercone and Luk [10] in 1988. The use of domain knowledge and the inferential structure of the data were highly emphasized in order to achieve a better quality performance. We are interested in adopting the advantages of semantic approaches on the null value problem to the processing of missing attribute values. Domain-related strategies is essential to preserve the original characteristics of the data set. Therefore semantic approaches of assigning missing values from

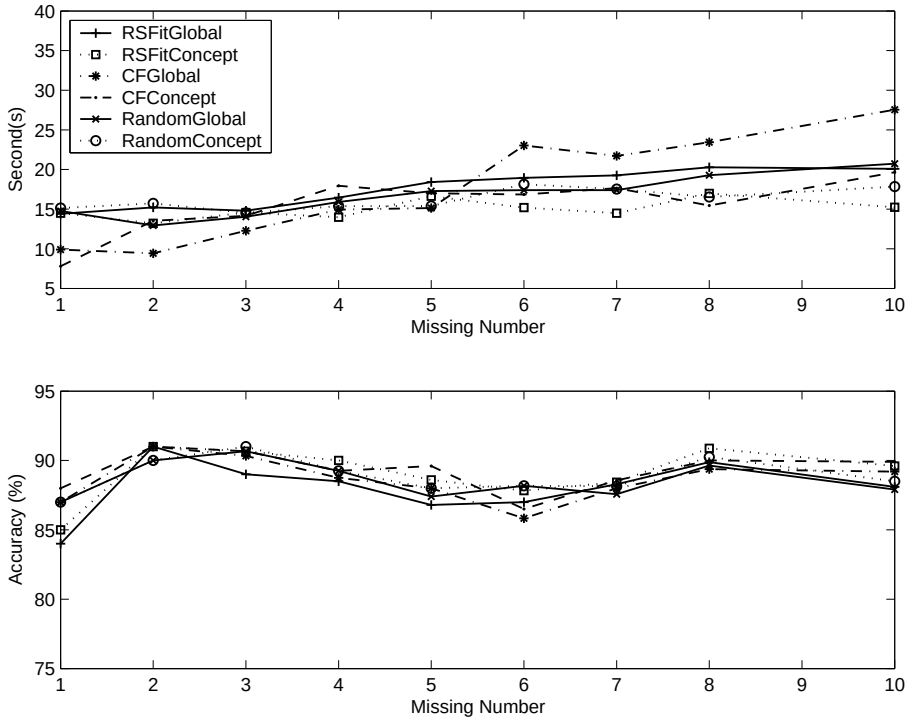

Fig. 4. Comparison Figure for Zoo Data

the domain knowledge may better keep the integrity of the data and provide a semantically precise data preprocessing.

Acknowledgements

We gratefully acknowledge the financial support of the Natural Science and Engineering Research Council of Canada. Thank Dr. Arnold Mitnitski from Department of Medicine, Dalhousie University for sharing the Geriatric Care data set.

References

1. Grzymala-Busse, J.W., Hu, M.: A Comparison of Serveral Approaches to Missing Attribute Values in Data Mining. W. Ziarko and Y.Yao (Eds.): RSCTC 2000, LNAI **2005** (2001) 378–385
2. Grzymala-Busse, J.W., Grzymala-Busse, W.J., Goodwin, L.K.: Coping with Missing Attribute Values Based on Closest Fit in Preterm Birth Data: A Rough Set Approach. Computation Intelligence. **17-3** (2001) 425–434

Table 5. Comparisons on Accuracies and Time For Geriatric Care Data

Data Sets	Computation Time (Second) For 100 Run					
Missing Values	RSFit Global	RSFit Concept	ClosestFit Global	ClosestFit Concept	Random Global	Random Concept
5	842.637	642.607	1673.972	1193.496	835.15	638.928
10	1660.891	1266.308	3337.450	2403.185	1645.864	1255.622
15	2481.925	1896.875	4993.163	3611.637	2454.688	1880.036
20	3298.103	2479.502	6668.741	4663.448	3265.128	2452.722
25	4118.382	3116.954	8315.181	5878.851	4106.38	3088.842
30	4933.933	3714.456	10126.725	7000.339	4928.184	3676.568
40	6595.240	4978.143	13375.369	9462.916	6552.399	4936.833
50	8183.797	6222.527	16557.562	11747.613	8188.923	6162.332
60	9908.241	7479.915	20024.138	14126.664	9807.790	7413.180
Data Sets	Average Accuracy (Percentage %) over 100 Times					
Missing Values	RSFit Global	RSFit Concept	ClosestFit Global	ClosestFit Concept	Random Global	Random Concept
5	73.6%	73.4%	72.4%	71.2%	69.2%	69.2%
10	72.6%	72.6%	73.5%	73.5%	71.1%	71.1%
15	72.6%	72.67%	73.13%	72.8%	70.67%	70.67%
20	72.05%	72.00%	74.10%	74.30%	70.85%	70.90%
25	72.56%	72.48%	74.00%	73.84%	70.60%	70.60%
30	73.37%	73.37%	74.40%	74.33%	71.50%	71.57%
40	72.15%	72.15%	73.23%	73.40%	70.28%	70.25%
50	73.78%	73.82%	73.86%	73.86%	72.06%	72.08%
60	72.43%	72.45%	73.38%	73.45%	71.35%	71.38%

Data Instances: 8535. Condition Attributes: 44.

3. Grzymala-Busse, J.W.: Incomplete Data and Generalization of Indiscernibility Relation, Definability and Approximations. In: Proceedings of RSFDGrC 2005, LNAI **3641** (2005) 244–253
4. Pawlak, Z.: Rough Sets. In Theoretical Aspects of Reasoning about Data. Kluwer, Netherlands (1991)
5. Hu, X., Lin, T., Han, J.: A New Rough Sets Model Based on Database Systems. Fundamenta Informaticae **59** no.2-3 (2004) 135–152
6. Aleksander Ohrn :Discernibility and Rough Sets in Medicine: Tools and Applications. PhD Thesis, Norwegian University of Science and Technology (1999)
7. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. (2000)
8. Newman, D.J., Hettich, S., Blake, C.L. and Merz, C.J.: UCI Repository of machine learning databases. University of California, Irvine, Department of Information and Computer Seiences (1998) http://www.ics.uci.edu/∼mlearn/MLRepository.html
9. Li, J. and Cercone, N.: Empirical Analysis on the Geriatric Care Data Set Using Rough Sets Theory. Technical Report, CS-2005-05, School of Computer Science, University of Waterloo (2005).
10. Kao, M., Cercone, N., Luk, W.-S.: Providing Quality Response with Natural Language Interfaces: The Null Value Problem. IEEE Transactions on Software Engi-

Table 6. Comparisons on Accuracies and Time For Spambase Data

Data Sets	Computation Time (Second) For 100 Run					
Missing Values	RSFit Global	RSFit Concept	ClosestFit Global	ClosestFit Concept	Random Global	Random Concept
5	343.581	262.547	1163.445	685.700	339.987	260.068
10	676.317	504.733	2310.732	1361.200	663.801	500.903
15	995.771	746.852	3458.372	2069.000	987.348	742.708
20	1323.940	986.152	4605.719	2667.500	1309.055	977.558
25	1647.742	1223.637	5752.945	3300.900	1629.186	1216.533
30	1970.233	1470.366	6896.710	3992.000	1949.636	1460.533
35	2299.640	1705.113	8051.766	4625.400	2270.247	1695.289
50	3276.769	2437.121	11642.691	5461.400	3236.639	2420.724
Data Sets	Average Accuracy (Percentage %) over 100 Times					
Missing Values	RSFit Global	RSFit Concept	ClosestFit Global	ClosestFit Concept	Random Global	Random Concept
5	78.20%	78.20%	79.40%	79.40%	76.60%	76.60%
10	77.80%	77.80%	80.60%	80.60%	77.00%	77.00%
15	76.67%	76.67%	80.13%	80.13%	74.13%	74.13%
20	73.85%	73.85%	75.65%	75.65%	70.75%	70.75%
25	76.48%	76.48%	78.76%	78.76%	74.44%	74.44%
30	76.53%	76.53%	79.60%	79.60%	74.33%	74.33%
35	75.86%	75.86%	78.74%	78.74%	73.74%	73.74%
50	77.80%	77.80%	79.78%	79.78%	74.50%	74.50%

Data Instances: 4601. Condition Attributes: 57.

neering **14-7** (1988) 959–984

Table 7. Comparisons on Accuracies and Time For Lymphography Data

Data Sets	Computation Time (Second) For 100 Run					
Missing Values	RSFit Global	RSFit Concept	ClosestFit Global	ClosestFit Concept	Random Global	Random Concept
1	14.269	13.913	8.063	6.760	12.93	11.717
2	13.532	13.275	10.124	15.255	15.561	13.208
3	15.454	14.292	11.765	14.185	16.332	14.781
4	16.92	14.237	15.006	17.814	18.189	14.566
5	17.965	15.09	16.964	17.351	17.926	15.36
6	18.273	17.511	26.036	20.546	20.259	17.352
7	20.626	17.38	28.503	20.468	20.331	16.582
8	21.842	17.418	30.979	20.712	21.264	18.651
10	24.121	19.558	34.579	23.815	24.405	19.444
Data Sets	Average Accuracy (Percentage %) over 100 Times					
Missing Values	RSFit Global	RSFit Concept	ClosestFit Global	ClosestFit Concept	Random Global	Random Concept
1	57.00%	57.00%	58.00%	62.00%	59.00%	62.00%
2	62.50%	64.50%	64.00%	65.00%	64.00%	64.00%
3	61.67%	61.00%	59.33%	61.67%	61.33%	60.00%
4	63.00%	63.00%	59.75%	60.75%	62.00%	61.75%
5	64.60%	65.60%	63.00%	63.00%	63.40%	63.80%
6	64.00%	64.50%	60.17%	60.83%	63.50%	63.33%
7	59.43%	61.14%	59.57%	60.28%	61.57%	61.28%
8	61.63%	62.88%	61.50%	61.63%	59.50%	60.50%
10	61.30%	62.50%	62.00%	62.70%	62.70%	63.30%

Data Instances: 148. Condition Attributes: 18.

Table 8. Comparisons on Accuracies and Time For Zoo Data

Data Sets	Computation Time (Second) For 100 Run					
Missing Values	RSFit Global	RSFit Concept	ClosestFit Global	ClosestFit Concept	Random Global	Random Concept
1	14.404	14.495	9.900	7.790	14.767	15.135
2	15.219	13.219	9.421	13.524	12.938	15.737
3	14.813	14.579	12.284	14.154	14.043	14.522
4	16.445	13.974	14.937	17.948	15.909	15.376
5	18.416	16.639	15.136	17.012	17.277	15.385
6	18.938	15.195	23.021	16.837	17.387	18.133
7	19.259	14.500	21.720	17.647	17.401	17.543
8	20.278	16.990	23.439	15.456	19.290	16.528
10	20.089	15.236	27.541	19.657	20.738	17.846
Data Sets	Average Accuracy (Percentage %) over 100 Times					
Missing Values	RSFit Global	RSFit Concept	ClosestFit Global	ClosestFit Concept	Random Global	Random Concept
1	84.00%	85.00%	87.00%	88.00%	87.00%	87.00%
2	91.00%	91.00%	91.00%	91.00%	90.00%	90.00%
3	89.00%	90.67%	90.33%	90.67%	90.67%	91.00%
4	88.50%	90.00%	88.75%	89.25%	89.25%	89.25%
5	86.80%	88.60%	87.99%	89.60%	87.40%	87.99%
6	87.00%	87.83%	85.83%	86.50%	88.17%	88.17%
7	88.29%	88.43%	88.00%	88.57%	87.57%	88.14%
8	89.88%	90.88%	89.38%	90.00%	89.63%	90.25%
10	88.10%	89.60%	89.20%	89.90%	87.90%	88.50%

Data Instances: 101. Condition Attributes: 16.