Addressing Missing Attributes During Data Mining Using Frequent Itemsets And Rough Set Based Predictions

Jiye Li
York University
jiye@cse.yorku.ca

Nick Cercone
York University
ncercone@yorku.ca

Robin Cohen
University of Waterloo
rcohen@uwaterloo.ca

Abstract

In this paper, we present an improved method for predicting missing attribute values in data sets. We make use of frequent itemsets, generated from the association rules algorithm, displaying the correlations between different items in a set of transactions. In particular, we consider a database as a set of transactions and each data instance as an itemset. Then frequent itemsets can be used as a knowledge base to predict missing attribute values. Our approach integrates the RSFit method based on rough sets theory that produces faster predictions by considering similarities of attribute value pairs, but only for those attributes contained in the core or reduct of the data set. Using empirical studies on UCI and other real world data sets, we demonstrate a significant increase in prediction accuracy obtained from our new integrated approach, referred to as ItemRSFit.

1 Introduction

How to process data containing missing attribute values is an important task in the data preprocessing stage for data mining applications. Missing attribute values commonly exist in real-world data sets. They may come from the collecting process, or from redundant diagnosis tests, change of the experimental design, privacy concerns, unknown data, etc. Various approaches on how to cope with the missing attribute values have been proposed in the past years. For example, in [5], proposed methods include "using the most common attribute value", "ignoring examples with unknown attribute values" and "assigning all possible values of the attribute restricted to the given concept". In [4], vectors of all attributes are compared and a distance function is used to determine the most similar attribute pairs, in order to assign the missing value. These approaches have difficulties, however. For example, selecting the most frequent value may lead to an inconsistent data set; assigning "unknown" to an attribute may not be reasonable in some instances (e.g. for gender, in a health data set).

We are interested in predicting missing attribute values in the data preprocessing stage of the knowledge discovery process. Motivated to develop a technique that can predict all the missing attribute values with high precision, we integrate two techniques into our solution.

The first is the association rule algorithm [1], which is well known in data mining for discovering item relationships from large transaction data sets. Prior to the association rule generation, frequent itemsets are generated based on the item-item relations from the large data set according to a certain **support**. Thus the frequent itemsets of a data set represent strong correlations between different items, and the itemsets represent probabilities for one or more items existing together in the current transaction. In our research, we are considering a certain data set as a set of transactions. The implications from frequent itemsets can be used to find which attribute value the missing attribute is strongly connected to and frequent itemsets can be used for predicting the missing values. We call this approach the "itemset-approach" for prediction. Apparently, the greater the number of frequent itemsets used for the prediction, the more information from the data set itself is used for prediction; hence, the higher the accuracy that will be obtained. However, generating frequent itemsets for a large data set is time-consuming. Lower support requirements lead to a larger number of frequent itemsets, but usually cost a significant amount of computation time. Although demanding higher support requires less computation time, the resulting itemsets show restricted item relationships and the applicable number of itemsets are fewer; therefore, not all the missing values can be predicted. In order to balance the tradeoff between computation time and the percentage of the applicable prediction, another approach must be taken into consideration.

The second element of our solution is rough sets theory, proposed in the 1980's by Pawlak [9] and used for attribute selection, rule discovery and many knowledge discovery applications in the areas such as data mining, machine learning and medical diagnoses. Core and reduct are

0-7695-3032-X/07 $25.00 © 2007 IEEE
DOI 10.1109/GrC.2007.144

among the most important concepts in this theory. The data set is viewed as a decision table where each data instance consists of condition attributes and decision attributes. A **reduct** contains a subset of condition attributes that are sufficient enough to represent the whole data set. The intersection of all the possible reducts is the **core**. Therefore the attributes contained in the reduct or core are more important and representative than the rest of the attributes. By examining only attributes within the same core or reduct to find the similar attribute value pairs for the data instance containing the missing attribute values, we can assign the most relevant value for the missing attribute. Since this method only considers a subset of the data set, which is either the core or the reduct, the prediction is faster than those considering the complete data set. This approach "RSFit" (Rough Set Fit) [7] is an alternative approach for fast prediction and it can be used to predict missing attributes that cannot be predicted by the frequent itemsets.

We integrate the prediction based on frequent itemsets and the RSFit approach into a new approach **ItemRSFit** (Itemset and Rough Set Fit) to predict missing attribute values. Experiments on UCI [3] data sets and a real world data set demonstrate our proposed approach on assigning missing attribute values obtains a high accuracy.

2 RSFit Approach to Assign Missing Values

In this section, we introduce the RSFit approach [7] for predicting missing values. We first make definitions to be used in the following descriptions of the proposed approaches. The input to our approach is a decision table $T = (C, D)$ containing missing attribute values, 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. Table 1 shows an example of a decision table, which may be used to decide the mileage category of a car, based on various features that describe the car.

The primary idea is to search only for attribute-value pairs within the core or the reduct of the data set. 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 only exist in the condition attributes not in the decision attributes.

First, we obtain the core of the data set $T = (C, D)$ based on Hu's core algorithm introduced in [6]. If the target attribute c_k does not belong to the core, we include c_k in the core. In case there is no core for T, we consider the reduct of T. ROSETTA software [8] is used for reduct generation. 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, $U' = \{u_1, u_2, \ldots, u_{n'}\}$, $1 \leq n' \leq n$. A distance function, such as Euclidean distance or Manhattan distance is then applied to compute the similarities between different attribute-value pairs. The best match has the smallest difference from the target attribute-value pair. When this is determined, we assign the value from the best matched attribute-value pair to the target missing value. Thus, a complete data set without missing values is obtained.

We demonstrate the RSFit approach by an artificial car data set which appeared in [6] as shown in Table 1. One missing attribute value for "compress" is randomly selected across the data set as shown by Table 2 a). First, the core is obtained for this data set as "Make_model" and "trans". Since the core attributes exist and the missing attribute "compress" does not belong to the core, we add attribute "compress" to the core set. The new data set containing the core attributes, target attribute "compress" and the decision attribute are created and shown in Table 2 b). Then we will find the match for attribute "compress" in u_8. u_{14} has the smallest difference, which is 0, from u_8, therefore, u_{14} is the best match. We assign $c_{compress_{14}}$ to $c_{compress_8}$, which is "high" (correct prediction).

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

3 ItemRSFit Approach

The RSFit approach cannot provide a very high prediction precision, although it is computationally faster than the "closest fit" approach of [4] (see [7]). This is because this approach does not fully consider the item-item relationships inside the data set. Take Table 1 as an example. The attribute value pairs such as $(make\_model_{usa}$, $displace_{medium}$, $compress_{high}$, $power_{high})$ frequently appear together. Such frequently appearing value pairs cannot be extracted by simply looking at the distance com-

Table 2. a) With One Missing Attribute Value, b) New Decision Table

U	...	compress	...
1	...	high	...
2	...	medium	...
3	...	high	...
4	...	medium	...
5	...	medium	...
6	...	medium	...
7	...	medium	...
8	...	?	...
9	...	high	...
10	...	medium	...
11	...	high	...
12	...	medium	...
13	...	high	...
14	...	high	...

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

parisons between different data instances by the RSFit approach. RSFit uses only a subset of the transaction set as a knowledge base to find the similar object for prediction.

The association rule algorithm was first introduced in [1], and it can be used to find associations among items from transactions. For example, in *market basket analysis*, by analyzing transaction records from the market (i.e. lists of items purchased, for each customer), we could use association rule algorithms to discover different shopping behaviours such as, when customers buy bread, they will probably buy milk.

Frequent itemsets generation is the first step for association rule generation. Itemsets that frequently occur together in the transactions are generated. Rules based on these itemsets are further extracted to represent the associated relations. Many contributions such as [2] on how to efficiently generate frequent itemsets and rules have been reported.

Our approach is to consider data in the form of a decision table as the transaction set for generating the frequent itemsets. Each attribute value is considered to be an item in the set of transactions. For example, in Table 1, (cyl_4, $door_2$), ($make\_model_{usa}$, $displace_{medium}$, $compress_{high}$, $power_{high}$) and ($power_{high}$, $trans_{auto}$, $weight_{medium}$) frequently occur together, and are considered as the frequent itemsets. We define the following concepts:

Definition *Transaction.* The set of transactions to the frequent itemsets generation is in a form of a decision table T=(C, D), where $C = \{c_1, c_2, \ldots, c_m\}$ is the condition attribute set where m is the number of condition attributes, and $D = \{d_1, d_2, \ldots, d_l\}$ is the decision attribute set where l is the number of decision attributes. $U = \{u_1, u_2, \ldots, u_n\}$ represents the itemsets in T, where n is the number of transactions in T. Each transaction contains $(m + l)$ items. Therefore each attribute value is considered an item in the transaction.

An association rule [1] is a rule of the form $\alpha \rightarrow \beta$. α and β represent itemsets which do not share common items. α contains items from the condition attribute set C, and β contains items from the decision attribute set D.

Definition *Support.* A support of an itemset is the percentage of the number of transactions containing the itemset to the total number of transactions. Support can be represented as $\frac{|\alpha \cup \beta|}{|T|}$. For example, in Table 1, it would be reasonable to extract an association rule where α_1 is $power_{high}$, $trans_{auto}$, $weight_{medium}$ and β_1 is $mileage_{medium}$. The support for $\{\alpha_1, \beta_1\}$ from Table 1 would be 21.4%.

Frequent itemsets generation in an association rule algorithm first counts the frequencies of each individual item among the whole transaction set. Then based on the 1-itemsets whose support are no less than the predefined minimum support, frequent 2-itemsets are generated. Those itemsets with occurrence no less than the minimum support are selected for frequent 3-itemsets generation. Frequent l-itemsets are generated based on the frequent $(l-1)$-itemsets. The process continues until no new frequent itemsets are found.

To predict missing attribute values, let $T = (C, D)$ be the decision table that contains missing attribute values, where $C = \{c_1, c_2, \ldots, c_k, \ldots, c_m\}$, $1 \leq k \leq m$, and $U = \{u_1, u_2, \ldots, u_n\}$, $1 \leq n$. First, the data input to the association rule algorithm is prepared. Data instances with missing attribute values are all removed from T, and we call the new decision table T''. T'' does not contain any missing values. Secondly, frequent l-itemsets are generated based on T'' with a given minimum support. Let $Itemsets = \{S_1, S_2, \ldots, S_g\}$, where S_i $(1 \leq i \leq g)$ is a frequent l-itemsets generated based on $T = (C, D)$ according to a minimum support, $S_i = \{v_{i1}, v_{i2}, \ldots, v_{il}\}$, l is the number of items contained in S_i, and v_{ij} $(1 \leq j \leq l)$ is an attribute value in T.

We use the frequent itemsets generated in the previous step as our knowledge base to find a match for the missing value. Let $u_i = (v_{1i}, v_{2i}, \ldots, v_{ki}, \ldots, d_i)$ $(1 \leq i \leq n)$ be the data instance in T containing the missing attribute value v_{ki} (represented as v_{ki} =?) for attribute c_k $(1 \leq k \leq m)$. We search from $Itemsets$ for all the itemsets containing the missing attribute v_k, and check which itemset among the itemsets can be **applied** to u_i. We say a frequent itemset can be applied to this data instance if all the items in this itemset, except the missing attribute, have exactly the same attribute values as contained by the data instance that has the missing attribute value. If this itemset can be applied, we assign the attribute value contained in this itemset to the missing attribute. 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.

Suppose u_i and u_j are two of the data instances in T

that contain missing attribute values. $u_i = (v_{1i} = 1, v_{2i} = 2, v_{3i} = 4, v_{4i} =?, v_{5i} = 1)$, and $u_j = (v_{1j} = 0, v_{2j} = 2, v_{3j} =?, v_{4j} = 5, v_{5j} = 0))$. The itemsets for T generated from the second step are $S_1 = \{v_1 = 1, v_3 = 4\}$, $S_2 = \{v_2 = 1, v_4 = 7, v_5 = 0\}$, and $S_3 = \{v_2 = 2, v_3 = 4, v_4 = 6, v_5 = 1\}$. For the missing value in u_i, S_1 cannot be used for predicting v_2 because it does not contain the missing attribute v_4. S_2 cannot be used for assigning the missing value either, because of the different values of v_2. Since all the items in S_3 can be applied to u_i, we assign $v_4 = 6$. For the missing value in u_j, none of the three itemsets can be applied to find a match. The missing value will be further processed by other missing attribute value processing approaches such as RSFit.

Missing attribute values from some data instances in the original data set can be predicted by frequent itemsets. We call these data instances **Compatible Records**.

Definition *Compatible Record.* A compatible record (CR) is a record whose missing attributes can be predicted by an itemset. More formally, a record r with p missing attributes is a CR if there exists an itemset I such that $|I \cap r| \leq p$. There also exist data instances for which no possible match can be found to predict the missing values. If a record is not CR, the RSFit method is applied to predict the rest of the missing attribute values. We call this integrated approach **ItemRSFit**. The details on the integrated approach is shown in the following Figure 1. **Stage A** illustrates the itemset approach, in which the frequent itemsets, as the knowledge base, are generated based on using the apriori association rule algorithm from complete data instances. The reduct and core are generated in **stage B** for the use of the RSFit approach. In **stage C**, the frequent itemsets are used to predict the missing attribute first, then the RSFit approach is applied to the rest of the missing cases.

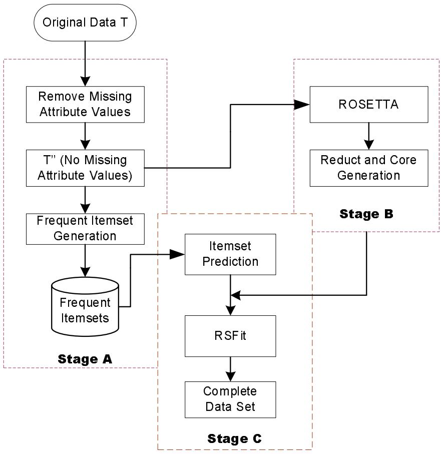
Original Data T
Remove Missing Attribute Values
T'' (No Missing Attribute Values)
Frequent Itemset Generation
Frequent Itemsets
Stage A
ROSETTA
Reduct and Core Generation
Stage B
Itemset Prediction
RSFit
Complete Data Set
Stage C

Figure 1. ItemRSFit Approach

4 Evaluation

We use the following approach to perform the evaluation process. We consider complete data sets as the transaction data set T. 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 then apply both the RSFit approach and the ItemRSFit approach on predicting missing values, and compare the accuracy of the predictions from these two approaches. For n missing attribute values, the **prediction accuracy** is defined to be the percentage of values that were correctly predicted (where the correct values are the ones removed to create the missing values for the experiment).

The ItemRSFit approach is implemented by Perl and the experiments are conducted on Sun Fire V880, four 900Mhz UltraSPARC III processors. We use apriori frequent itemsets generation [2] to generate frequent 5-itemsets. The core generation in the RSFit approach is implemented with Perl combining the SQL queries accessing MySQL (version 4.0.12). ROSETTA software [8] is used for reduct generation.

4.1 Experiments on Geriatric Care Data

We first perform experiments on a geriatric care data set. This data set is an actual data set from Dalhousie University Faculty of Medicine to determine the survival status of a patient giving all the symptoms he or she shows. The data set contains $8,547$ patient records with 44 symptoms and their survival status. We use *survival status* as the decision attribute, and the 44 symptoms of a patient as condition attributes, which includes *education level, the eyesight, live alone, cough, high blood pressure, heart problem, gender, the age of the patient at investigation* and so on. 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 $8,535$ records. [1] The core attributes for this data set are *eartrouble, livealone, heart, highbloodpressure, eyetrouble, hearing, sex, health, educationlevel, chest, housework, diabetes, dental, studyage*.

ItemRSFit approach is the new integrated approach introduced in this paper. Table 3 lists the prediction accuracy for both RSFit and ItemRSFit. We also list the number and the percentage of CR by only using frequent itemsets as knowledge for prediction. In this research, we experiment on geriatric care with 50 to 200 missing attribute values.

From Table 3 we can see, the smaller the support becomes, the more itemsets are generated and the larger the

[1] Note that the core generation algorithm cannot return correct core attributes when the data set contains inconsistent data entries.

Table 3. Comparisons for Geriatric Care Data on Prediction Accuracy

Data Sets	Average Accuracy(Percentage%)				
Missing Values	RSFit	Support	# CR	% CR	Integrated ItemRSFit
50	64.00%	90%	11	22%	64.00%
		80%	22	44%	68.00%
		70%	26	52%	68.00%
		60%	38	76%	72.00%
		50%	41	82%	70.00%
		40%	43	86%	72.00%
		30%	43	86%	78.00%
		20%	46	92%	90.00%
		10%	46	92%	96.00%
100	69.00%	90%	26	26%	69.00%
		80%	53	53%	74.00%
		70%	58	58%	74.00%
		60%	69	69%	77.00%
		50%	80	80%	75.00%
		40%	87	87%	76.00%
		30%	87	87%	81.00%
		20%	95	95%	87.00%
		10%	95	95%	96.00%
150	73.33%	90%	43	29%	75.33%
		80%	85	57%	79.33%
		70%	94	63%	79.33%
		60%	120	80%	80.00%
		50%	133	89%	81.33%
		40%	137	91%	82.00%
		30%	137	91%	83.33%
		20%	142	95%	89.33%
		10%	142	95%	96.67%
200	73.50%	90%	39	20%	73.50%
		80%	103	52%	77.00%
		70%	118	59%	76.50%
		60%	146	73%	75.50%
		50%	169	84%	73.50%
		40%	182	91%	79.00%
		30%	182	91%	79.50%
		20%	192	96%	88.50%
		10%	194	96%	96.00%

number of compatible records from frequent itemset becomes. ItemRSFit approach always maintains or improves on the prediction accuracy of RSFit.

Figure 2 shows the comparison for the number of CR by Itemsets prediction according to different support for different number of missing values. Frequent itemsets with lower support value can provide a larger knowledge base to find predictions. We can also see from Figure 2 that using itemsets alone cannot predict all the missing values. For instance, when there are 50 missing values existing in the data set, given $support = 10\%$, there are still 8% missing instances that cannot be predicted by the itemsets.

In order to show that ItemRSFit approach obtains better prediction accuracy than RSFit, we show the prediction accuracy comparisons on geriatric care data set with 150 missing attribute values, as shown in Figure 3. We can see when support value is lower, the prediction accuracy of ItemRSFit is significantly higher than RSFit.

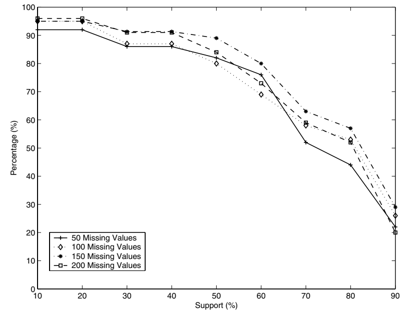

Figure 2. Comparisons on the Percentage of Compatible Records for Geriatric Care Data using Frequent Itemsets to Predict

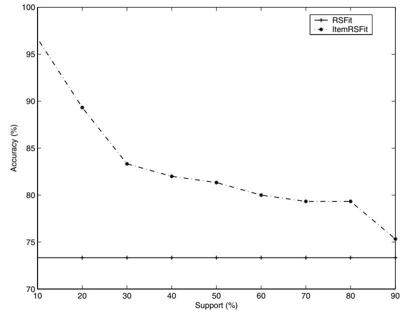

Figure 3. Accuracy Comparisons for Geriatric Care Data with 150 Missing Attribute Values

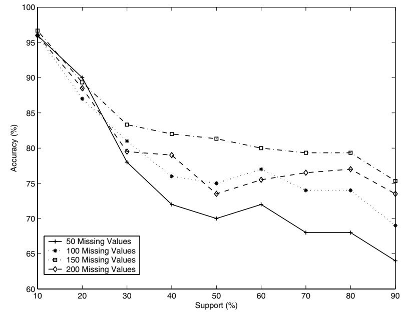

Figure 4. Accuracy Comparisons for Geriatric Care Data with Different Number of Missing Attribute Values

Figure 4 demonstrates the prediction accuracy comparisons for different number of missing attribute values with different support for the geriatric care data set using ItemRSFit. The ItemRSFit approach obtains higher accuracy when the support value is lower and this is unaffected by the number of missing attribute values in the data set.

Observations. From the experimental results on the geriatric care data set shown in Figures 2, 3, 4, we observe the prediction accuracy for the ItemRSFit approach increases while the support value decreases. The frequent Itemsets approach can provide a higher prediction by itself. But this approach cannot predict all the missing values in the geriatric care data set. For the ItemRSFit approach on geriatric care data, the highest accuracy is obtained when $support = 10\%$; the lowest accuracy is obtained when $support = 90\%$. This can be explained as follows. "Support" is a measure to evaluate the occurrence of both the antecedents and the consequents of an association rule in the data set. The higher the support is, the more frequent this occurrence has to be and the less knowledge for prediction is obtained. When the support value is increased, fewer matched cases are found from the itemset approach; therefore, more missing values have to be predicted by the RSFit approach. The lowest accuracy of the ItemRSFit approach is equal to the accuracy from the RSFit approach. The RSFit approach gives the baseline prediction accuracy for the ItemRSFit approach. For different numbers of missing attribute values, the frequent itemsets with the lowest support brings the highest prediction accuracy. The frequent itemsets alone as the knowledge base to predict the missing values cannot fully find all the matches for the missing value for geriatric care data.

4.2 Experiments on UCI Data Sets

In the experiments on the UCI data sets [3] we study how the ItemRSFit approach can be applied for predictions on different types of data sets. We experiment on data sets with no missing attribute values. For data sets with continuous attributes, we discretize the attribute values to discrete data before generating frequent itemsets. For each data set, we randomly select a certain percentage of the total possible missing values (total number of condition attributes × total number of data instances) as missing attribute values, and list the prediction accuracy comparisons for the ItemRSFit and RSFit approaches according to different support values.

Experimental results from UCI abalone, glass, iris and lymphography data sets [3] have demonstrated the high prediction characteristics of the proposed ItemRSFit approach on processing data with missing attribute values as shown in Figure 5-8. Frequent itemsets can be used as a knowledge base to predict missing attribute values.

4.3 Discussions and Related Work

Experimental results from both the real-world geriatric care data set and UCI data sets have demonstrated the high prediction characteristics of the proposed ItemRSFit approach on processing data with missing attribute values. The frequent itemsets can be used as a knowledge base to predict missing attribute values.

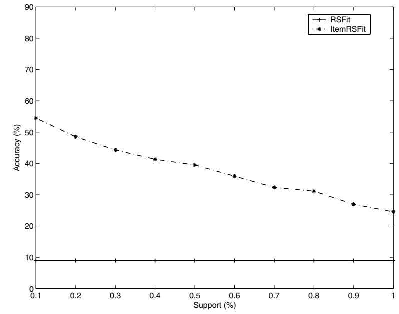

Figure 5. Abalone Data with 0.5% Missing Attribute Values

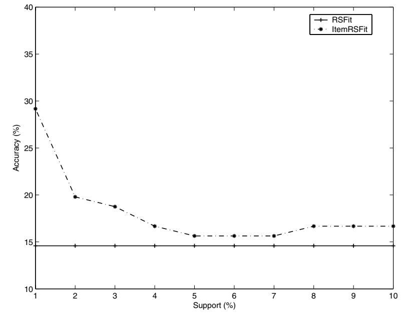

Figure 6. Glass Data with 5% Missing Attribute Values

We find the approach introduced in [10] close to our work. An approach of using association rules generations on completing missing values is discussed. However, our proposed ItemRSFit approach is quite different. First, only frequent 1-itemsets and 2-itemsets are used in [10] to find the possible values for the missing data, and data associations with missing attributes on the consequent part are used for prediction. What percentage of the missing data can be predicted with the data association is not discussed. We use frequent 5-itemsets as the knowledge base for prediction as described in the experimental design part in Section 4. We explore the relations between different support and the percentage of the **compatible records** using frequent itemsets as shown in Figure 2. Second, in case there is no match from the data association, the missing value is assigned by the most common value of the missing attribute in [10]. We use frequent itemsets as the knowledge base for prediction, and the RSFit approach for the **non-compatible records** where the itemset cannot be applied, which guarantee that more important attributes are taken into considerations while predicting attributes. The proposed ItemRSFit approach provides predictions based on the data domain itself, which better preserves the originality of the data sets and avoids noise. Third, in [10], data associations, which are similar to associated rules, are generated according to

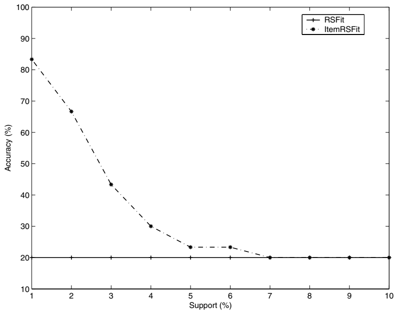

Figure 7. Iris Data with 5% Missing Attribute Values

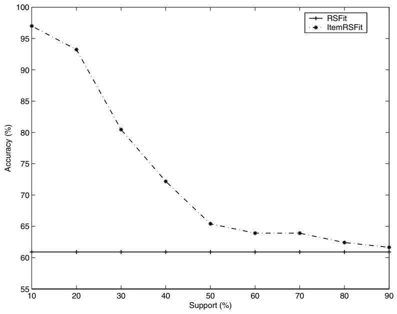

Figure 8. Lymphography Data with 5% Missing Attribute Values

both support and confidence and used as a knowledge base for predictions. Our approach is more efficient because we do not need to generate associated rules based on both support and confidence for prediction. Only support is used for frequent itemsets generation in the ItemRSFit approach.

5 Concluding Remarks and Future Work

We explore a new usage of association rule algorithms to predict missing attribute values, combined with rough sets theory. We first introduce a new approach RSFit to assign missing attribute values based on rough sets theory. Comparing to the "closest fit" approach [4], this approach significantly reduces the computation time and a comparable accuracy is achieved. We then introduce an integrated approach ItemRSFit based on both the association rule algorithm and rough sets theory to assign missing attribute values. The experimental results show the new approach obtains high prediction accuracy. It relies on its own data as a knowledge base and therefore the predicted values are not biased.

In our research, we adopt the strategies used by [11] on balancing the computational cost and the prediction accuracy. Lower support value can bring a higher prediction accuracy; however, frequent itemsets with lower support requires more time for computation than frequent itemsets with higher support. In the future, we are interested in exploring a satisfactory balance between the support value and the prediction accuracy. Given the available computational cost and the affordable computation time, it is interesting to explore what percentage of the missing attributes can be effectively predicted, and what are the most effective attributes to be predicted. In case of a higher prediction cost, the idea of giving more important attributes higher priorities for predictions can be applied as an heuristic.

Acknowledgements

Thanks to NSERC for financial support and to Dr. A. Mitniski of Dalhousie University for the Geriatric Care dataset.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499, 1994.

[2] C. Borgelt. Efficient implementations of apriori and eclat. In *Proceedings of the FIMI'03 Workshop on Frequent Itemset Mining Implementations, Melbourne, Florida, USA, November, CEUR Workshop Proceedings*, 2003.

[3] C. B. D.J. Newman, S. Hettich and C. Merz. UCI repository of machine learning databases, 1998.

[4] J. W. Grzymala-Busse, W. J. Grzymala-Busse, and L. K. Goodwin. Coping with missing attribute values based on closest fit in preterm birth data: A rough set approach. *Computational Intelligence*, 17(3):425–434, 2001.

[5] J. W. Grzymala-Busse and M. Hu. A comparison of several approaches to missing attribute values in data mining. In *Rough Sets and Current Trends in Computing*, pages 378–385, 2000.

[6] X. Hu, T. Y. Lin, and J. Han. A new rough sets model based on database systems. *Fundam. Inf.*, 59(2-3):135–152, 2004.

[7] J. Li and N. Cercone. Assigning missing attribute values based on rough sets theory. In *IEEE Granular Computing*, pages 607–610, 2006.

[8] A. Øhrn. *Discernibility and Rough Sets in Medicine: Tools and Applications.* PhD thesis, Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim Norway, 1999.

[9] Z. Pawlak. *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Norwell, MA, USA, 1992.

[10] C.-H. Wu, C.-H. Wun, and H.-J. Chou. Using association rules for completing missing data. In *Fourth International Conference on Hybrid Intelligent Systems*, pages 236–241, 2004.

[11] X. Zhu and X. Wu. Cost-constrained data acquisition for intelligent data preparation. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1542–1556, 2005.