Rough Set-Based Reduction of Incomplete Medical Datasets by Reducing the Number of Missing Values

Luai Al Shalabi
Faculty of Computer Studies, Arab Open University, Kuwait

Abstract: *This paper proposes a model of: firstly, dimensionality reduction of noisy medical datasets that based on minimizing the number of missing values, which achieved by cutting the original dateset, secondly, high quality of generated reduct. The original dataset was split into two subsets; the first one contains complete records and the other one contains imputed records that previously have missing values. The reducts of the two subsets based on rough set theory are merged. The reduct of the merged attributes was constructed and tested using Rule Based and Decomposition Tree classifiers. Hepdata dataset, which has 59% of its tuples with one or more missing values, is mainly used throughout this article. The proposed algorithm performs effectively and the results are as expected. The dimension of the reduct generated by the Proposed Model (PM) is decreased by 10% comparing to the Rough Set Model (RSM). The proposed model was tested against different medical incomplete datasets. Significant and insignificant difference between RSM and PM are shown in Tables 1-5.*

Keywords: *Data mining, rough set theory, missing values, reduct.*

Received October 9, 2015; accepted August 24, 2016

1. Introduction

Data mining is the process for dealing with problems such as classification, association, and prediction. The goal of data mining process is to predict a data mining model with some certainty while analyzing a small specific representative part of the data. Rough set theory, Rule based and decomposition tree are some examples of classifiers that discover knowledge. Two different challenges in data mining are consider here: finding the minimal reduct and solving incomplete data. In the next subsections, a review of each of the two challenges will be viewd.

Accuracy and efficiency are two important ways that are used to test a data mining model. One way to improve the accuracy and efficiency of a huge dataset is to distribute it into different subsets based on some criteria. There are different applications to do that including MapReduce and HBase [19] and cloud computing [18].

This work first formed a model in order to minimize the number of incomplete data in medical datasets. Then, it built the minimal reduct from the model (which is the dataset of minimal number of incomplete data). The generated minimal reduct, when tested; it gave the high accuracy results.

This work can be applied to all datasets that can be logically distributed vertically into two or more portions. Medical datasets are just a random choice of area. Each dataset can be logically distributed into two main portions: clinical and pathological.

1.1. Minimal Reduct

One common challenge in the data mining field is to find the minimal reduct which has a quality rate similar to the original dataset. According to Al Shalabi *et al*. [2], the data in the information system can be used to discern classes only to a certain degree. Not all attributes may be required in order to be able to do so. Therefore, discovering dependencies between attributes enables the reduction of the set of attributes.

Rough set theory which was introduced by Pawlak and Skowron [15] is an important theory for classification problems. The theory is more powerful in solving problems of data reduction, discovery of data dependencies, and dealing with missing values.

Reduction of attributes has been carried out by many researchers. A method to find the minimal reduction was proposed by Ye *et al*. [21]. Reduction of attributes under variable-precision dominance-based rough sets was investigated by Inuiguchi *et al*. [9]. A minimum cost attribute reduction was proposed by Jia *et al*. [11]. A study of attribute reduction in inconsistent incomplete decisions systems was done by Meng *et al*. [12]. A new algorithm for finding the reduction of attributes based on fuzzy rough sets was proposed by Chen *et al*. [6]. A distance measure approach to explore the boundary region for attribute reduction was proposed by Parthalain *et al*. [14]. The attribute reduction based on new conditional entropy for incomplete decision systems was studies by Dai *et al*. [7]. Jacob and Raju [10] investigated the preeminent feature selection and prediction technique that enhanced the software fault prediction accuracy with the optimal set of features.

1.2. Incomplete Data

Another challenge in the data mining field is incomplete data [3]. Missing Values (MV) usually affect the accuracy of a data mining system. It has been stated in many studies that the representation of imputed dataset may no longer be good and it may lead to the solutions that are far from optimal [5]. Reducing number of missing values in the dataset is a good solution to overcome this problem. It would improve the representation of the dataset by decreasing the error rate of bad imputation. Medical datasets should be treated carefully and this solution would help.

Missing data might occur because the value is not relevant to a particular case, could not be recorded when the data was collected, or is ignored by users because of privacy concerns [1]. The problem of missing values and solutions for this problem has been investigated long time ago [16]. The difficult problem is how to guess the suitable missing values. The literature review indicates that imputation practices are varied, and the evaluation of imputation accuracy is still a challenge [23]. Dempster *et al*. [8] proposed methods for solving missing data in datasets. They summarized these methods in three categories. First: Ignoring and discarding data by deleting all records that have missing data. Second: Parameter estimation which is used to find the parameters for the complete data. This method is use the Expectation-maximization algorithm for handling the parameter estimation of the missing data. Third: Imputation technique which replaces the missing values based on estimated values that are the most probable values. Imputation is a method that involves replacing an incomplete observation with complete information based on an estimate of the true value of the unobserved variable [13]. Zhang [22] introduced a new imputation approach called Shell Neighbors Imputation (SNI). The SNI fills in an incomplete instance in a given dataset by only using its left and right nearest neighbors with respect to each factor, referred them to Shell Neighbors.

2. Materials and Methods

The PM was tested against different incomplete medical datasets. Hepdata dataset is one of them. The detailed steps of applying the PM to Hepdat dataset will be explained throughout this article. In this research paper, different techniques and methods will be used as follows:

1. Methods for imputing missing values. The average value per class and the most probable value per class were chosen to impute missing values based on the type of the variable (continuous or categorical).
2. Technique for finding the reduct of the dataset. Rough set theory will be used for this purpose. It will be explained in the next section.
3. Classifiers for testing the validity of the generated reduct which producing accuracy, coverage, and number of nodes.
4. These classifiers are:

a) Rule Based Classifier (RBC).
b) Decomposition Tree Classifier (DTreeC).

Rule based classifier has to come up with a model from the dataset. The dataset contains one or more classification attributes while the remaining attributes are the conditional attributes. The model consists of the rules that govern classification.

Decomposition trees are used to split dataset into fragments not larger than a predefined size. These fragments, after decomposition represented as leafs in decomposition tree, are supposed to be more uniform and easier to cope with decision-wise. Usually the subsets of data in the leaves of decomposition tree are used for calculation of decision rules [4].

3. Rough Set Theory

Rough set theory is a mathematical tool to deal with uncertainty [14]. It can provide a tool for discovering relationships between records and decisions. So, the dataset can be reduced to get the minimum representation in terms of decision. The main basic concepts of rough set theory as explained by Pawlak are introduced.

3.1. Approximation of Sets

The lower approximation of *X* is the collection of objects that can be classified with full certainty as members of the set *X* using the attribute set A. Similarly, the upper approximation of *X* is also the collection of objects that possibly are classified as members of the set *X*. The boundary region comprises the objects that cannot be classified with certainty to be neither inside *X* nor outside *X*, using the same attribute set *A*.

Let $P \subseteq A$ and $Y \subseteq U$. The P-lower approximation of Y, denoted by $\underline{P}Y$, and P-upper approximation of *Y*, denoted by $\overline{P}Y$, are defined as:

$$\underline{P}Y = \cup\{X \in U|P: X \subseteq Y\} \tag{1}$$

$$\overline{P}Y = \cup\{X \in U|P: X \cap Y = \emptyset\}. \tag{2}$$

The *P*-boundary of set *Y* is defined as:

$$Bnp(y) = \overline{P}Y - \underline{P}Y \tag{3}$$

Set $\underline{P}Y$ is the set of all objects from *U* which can be certainly classified as elements of *Y*, employing the set of attributes *P*. Set $\overline{P}Y$ is the set of objects from *U* which can be possibly classified as elements of *Y*, using the set of attributes *P*. The set Bnp (*Y*) is the set of objects which cannot be certainly classified as element of *Y* using the set of attributes *P* only.

3.2. Rough Classification

Let S be an information system, P $\subseteq$ A, and let *Y*= {*Y1*, *Y2*, …, *Yn*} be a partition of *U*. By the P-lower approximation of *Y* in S it means the sets

$$\underline{P}Y = (\underline{P}Y1, \underline{P}Y2, \ldots, \underline{P}Yn\}. \tag{4}$$

The coefficient:

$$\gamma\ p(Y) = \sum_{i=1}^{n} card(\underline{P}Y) / card(U) \tag{5}$$

is called the quality of classification. It expresses the ratio of all P-correctly classified objects to all objects in the system.

3.3. Reduction and Dependency of Attributes

In the information system S, the minimal subset R $\subseteq$ P $\subseteq$ A such that $\gamma P(Y) = \gamma R(Y)$ is called Y-reduct of P and denoted by REDY (P). Notice that an information system may have more than one Y-reduct. Intersection of all Y-reducts is called the Y-core of P denoted by COREY (P), i.e., COREY(P) = $\cap$ REDY(P).

Discovering dependencies among attributes is the primary importance in the rough set approach to knowledge analysis. Set of attributes Q $\subseteq$ A depends on set of attributes P $\subseteq$ A, denoted by P $\to$ Q, if each equivalence class of the equivalence relation generated by P is included in some equivalence class generated by Q, i.e.

$$P \to Q \text{ iff } IND(P) \subseteq IND(Q). \tag{6}$$

4. The Proposed Work

4.1. The Model

A distributional model was proposed to find the minimal reduct from incomplete medical datasets. The proposed model first treats the missing values in datasets. Imputing missing values may damage the representation of the dataset if the guessing values are not correct. To cope with this problem, the PM minimizes number of missing values in the dataset by removing the columns with low coefficient which represents the weak relationship with other columns in the dataset. Consequently, number of imputed missing values in the dataset will be minimized. The incorrect imputation in the dataset will then be decreased and the error rate of bad guessing will be also decreased. The execution time comparison is not considered in this research since the main focus is on the minimizing of the number of missing values in the dataset. Execution time comparison will be studied and analyzed in the further research.

The PM works as follows: The imputation process was applied to the whole Original Data Set (ODS). Imputations by the average per class and by the most probable value per class which refers to the most repetitive value in the same column of the same class were used. After that, splitting the imputed dataset into two subsets was taken place. DS1 is the dataset of tuples without missing values and DS2 is the dataset of tuples with imputed values. The reduct was generated for DS1 and DS2 separately. It is not preferred to ignore and delete tuples of missing values (DS2) because it could add value to the final reduct since it has an amount of knowledge that cannot be ignored. The two reducts of DS1 and DS2 were merged in MRDS1RDS2 dataset. The reduct of MRDS1RDS2 was constructed and the resulted dataset (NewRDS) has lower number of imputed values. If some of these imputed values were wrongly guessed during the first stage, the representation of the NewRDS would not be highly affected because they are few.

In order to compare the proposed model with the rough set model, the steps that describe rough set model which is used in this study were listed as follows:

1. Obtain the original dataset with missing values. The proportions of missing values are shown in Table 1.
2. Fill in missing values using average per class and mode per class methods.
3. Find the reduct of the dataset after it is imputed using the reduction concept of rough set theory.
4. Train the dataset using the DTreeC and RBC.
5. Generate a model (a classification model).

Figure 1 describes the RSM and Figure 2 describes the PM. The comparison between the RSM and the PM was made and it will be discussed in the next section.

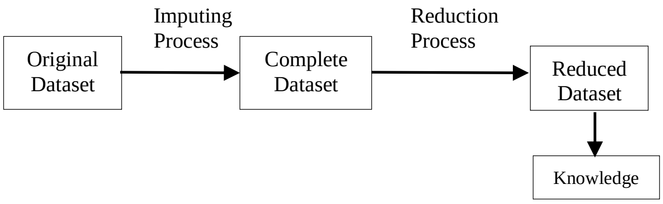

Figure 1. The rough set model (RSM).

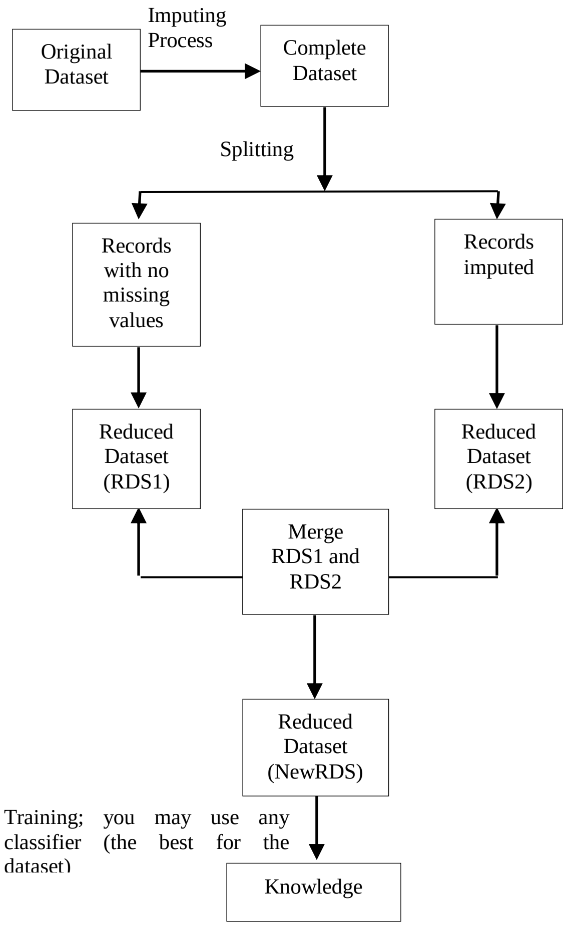

Figure 2. The proposed model (PM).

The algorithm called Reducing Missing Values (RMV) was constructed from the PM as shown below.

Algorithm 1: RMV algorithm

```
do for each column in the ODS
    do for each class
        if (the variable is continuous)
            CA ← the average value per class // CA is the calculated value
        else //categorical
            CA ← the most probable value per class
        do for each row in the class
            MV ← CA
        end do
    end do
end do
$DS1 \leftarrow Select_{(All\ records\ without\ missing\ values)}(ODS)$
$DS2 \leftarrow ODS - DS1$
$RDS1 \leftarrow RED(DS1)$
$RDS2 \leftarrow RED(DS2)$
$MRDS1RDS2 \leftarrow RDS1\ U\ RDS2$
$NewRDS \leftarrow RED(MRDS1RDS2)$
$Knowledge \leftarrow Train(NewRDS)$
    a. $(Knowledge)_{RBC} \leftarrow Train_{RBC}(NewRDS)$
    b. $(Knowledge)_{DTreeC} \leftarrow Train_{DTreeC}(NewRDS)$
```

4.2. Practical Example

Hepdata, Dermatology, Heart-Disease, and HSV datasets, available from the UCI Machine Learning repository, were employed for this study [20]. Hepdata (as an example) has 468 records. DS1consists of 192 records and DS2 consists of 276 records. The reducts of DS1 and DS2 were constructed. The conditional attributes that compose the reduct of DS1 are: 1,3,6,9,10,12,20 besides the class attribute. The conditional attributes that compose the reduct of DS2 are: 1,2,6,7,9,10,15,19 besides the class attribute. RDS1 and RDS2 are the datasets that were designed based on the reducts of DS1 and DS2 respectively. Both reducts (RDS1 and RDS2) were merged in MRDS1RDS2. The reduct of MRDS1RDS2 was constructed. It consists of the following nine conditional attributes: 1,2,6,7,10,12,15,19,20. These attributes represent the minimal reduct. The resulted dataset was trained using the RBC and DTreeC techniques in order to test the correctness of the generated reduct. The accuracy, the coverage percentage and/or the number of nodes were calculated as needed.

The other datasets were employed in the same way and the results were shown in Tables 1, 2, and 3. Table 1 shows the comparison between the PM and the RSM based on the cardinality of the minimal reduct. Table 2 shows the comparison between the PM and the RSM based on the number of clinical and histopathological attributes in each reduct. Less number of histopathological attributes is one of the targets since it minimizes the cost (money and moral). Table 3 shows the accuracy comparison based on the reducts generated by the PM and the RSM using RBC and DTreeC.

5. Results

5.1. Analysis of Results

The PM and the RSM were tested against the Dermatology, Heart-Disease, and HSV datasets as well as the Hepdata dataset. The datasets were chosen to cover different sizes based on the number of columns in the dataset varying from 12 to 35 columns. Also, they were chosen to have different proportions of tuples of missing values ranging from 15% to 96.6%.

Table 1 compares number of attributes (cardinality of the reduct) between the RBM and the PM after they were applied to the mentioned datasets. The difference between the sizes of both reducts generated by the RSM and the PM were calculated as below:

$$\textit{The difference} = \textit{Size(rough set reduct)} – \textit{Size(proposed reduct)} \quad (7)$$

If the value is positive (+) then the PM has an advantage over the RSM. If the value is negative (-) then the RSM has an advantage over the PM. Otherwise, both models have the same size of the reduct.

The percentage of the difference was also calculated and given by:

$$\textit{The percentage} = \textit{The difference/Size(rough set reduct)} \quad (8)$$

The advantage or disadvantage of the PM is given by the percentage value. If the value is positive (+) then the PM has an advantage over the RSM. If the value is negative (-) then the RSM has an advantage over the PM. Otherwise, both models are equal

Table 1. Comparisons based on cardinalities of the reducts.

Dataset	**the proportions of tuples of missing values**	**Number of conditional attributes in the reduct**		**The advantage of the PM over the RSM**	
		RSM	**PM**	**The difference in the two reduct's sizes**	**The percentage ~ (%)**
Hepdata	59%	10	9	+1	+10
Dermatology	15%	7	7	0	0
Heart-Disease	96.6%	3	2	+1	+33.3
HSV	25.4%	2	2	0	0

In Table 1, the reducts resulted from Hepdata and Heart-Disease datasets by the PM are significant since both consist of less cardinality comparing to the RSM whereas they consist of the same cardinality for Dermatology and HSV datasets. The importance of this study concluded by the advantages of having small cardinality (minimal reduct) and these advantages are, but not limited to the followings:

1. Saving the training and diagnosing times.
2. Reduce the cost of lab tests which could strain the patient (Moral and Monetary).

The reducts of Hepdata and Heart-Disease datasets generated by the PM have the advantages over the RSM since each of which has lower cardinality and this will save doctors' and patients' time as well as the patients' money and moral. Figure 3 illustrates this comparison.

Table 2 summarizes the reducts that were generated by both models. Number of clinical attributes and number of laboratory attributes in each reduct were counted. Laboratory attributes are complicated and costlier comparing to the clinical attributes that the consultant knows based on his/her previous experience and study. Lab tests are needed to take place in order to know the laboratory attributes' results. Minimizing the laboratory attributes in the reduct is one of the targets. Table 2 shows significant results that the PM generated less number of laboratory attributes comparing to the RSM and it is the case for all datasets under study.

The comparison between RSM and PM based on the number of missing values in each dataset was made. The decreasing percentage of missing values was also shown as in Table 3. The results are significant and they showed that the decreasing percentage of missing values in each reduct generated by the proposed model is very high comparing to the reducts generated by the rough set model except for the HSV reduct where both models gave closer percentage values.

Table 2. Reducts and analysis.

Dataset	**Rough set Model (RSM)**			**Rough set-based Proposed model (PM)**		
	Reduct	**# of CAt**	**# of LAt**	**Reduct**	**# of CAt**	**# of LAt**
Hepdata	1,2,6,7,**8,9**,10,15,**18**,20	0	11	1,2,6,7,10,**12,**15,**19**,20	0	10
Dermatology	3,**4**,17,22,**27,31**,33	2	5	3,**6,16**,17,22,33,**34**	3	4
Heart-Disease	**3**,5,**12**	1	2	**1**,5	1	1
HSV	**5**,10	0	2	**2**,10	1	1

Rough Set Exploration System (RSES) was used as a tool to generate the accuracy of the dataset. It is a toolset for analyzing data with the use of methods coming from Rough Set Theory. It is a graphical, user-friendly front-end running under Windows NT/98/95/2000/XP and providing access to methods from RSESlib library. RSESlib is a core of RSES' computational kernel. The RSES GUI allows point-and-click operation for making Rough Set computations. Both library and GUI are designed and implemented at the Group of Logic, Institute of Mathematics, Warsaw University and the Group of Computer Science, Institute of Mathematics, University of Rzeszów, Poland [17]. The comparison based on accuracy was established between RBM and the PM. This comparison represents the quality of the reduct.

Rule based classifier and decision tree classifier are the two classifiers considered to compare the accuracy of the two reducts: reduced information system. Two different reducts were generated for each dataset; one reduct generated by rough set theory reduction concepts and the other reduct was generated by the proposed method which employed rough set reduction concepts.

Table 3. Comparison between RSM and PM based on number of missing values in each reduct and the percentage decreased.

Dataset name	**Original data**	**Reduct of rough set (RSM)**		**Reduct of proposed model (PM)**	
	Number of missing values	**Number of missing values**	**The percentage decreased (%)**	**Number of missing values**	**The percentage decreased (%)**
Hepdata	167	124	**25.8**	41	**75.5**
Dermatology	301	69	**68.1**	45	**85.1**
Heart-Disease	161	114	**29.2**	6	**96.3**
HSV	96	15	**84.4**	20	**79.2**

Table 4. Comparisons based on accuracy tested by RBC and DTreeC.

	RBC (%)		**DTreeC (%)**	
Dataset	**Rough set reduct**	**Proposed reduct**	**Rough set reduct**	**Proposed reduct**
Hepdata	99.6	99.6	99.5	99.5
Dermatology	86.4	97.3	87.7	97.2
HeartDisease	100	100	100	100
HSV	100	100	100	100

As shown in Table 4, the accuracy that was given by RBC and DTreeC when they were applied to Dermatology dataset using the PM is significant and it was higher than that of the RSM whereas it is the same for the other three datasets in both RBM and PM.

Based on the main goal of this study, the PM always achieves the minimal reduct with the same or higher accuracy than the RBM. When the PM gave less cardinality, the accuracy was high and it was the same in both models. When the PM gave the same cardinality, the accuracy was same in both models as in HSV dataset and it was higher as in Dermatology dataset. For dermatology dataset, the PM still generates better reduct than the RSM (higher quality of reduct with same cardinality but higher accuracy). For HSV dataset where both models have the same accuracy given by RBC and DTreeC, other features such as coverage percentage and number of nodes were taken into account in order to decide the best model of higher quality.

DTreeC generates a tree with nodes that allow the classifier to classify a new object. The performance of a tree is measured by many features including number of nodes in the tree. When the DTreeC was applied to the dataset generated by the PM (reduct), number of nodes generated from the HSV dataset was 3, whereas it was 7 when the same classifier was applied against the dataset generated by the RSM (the reduct). The size of the tree based on the reduct generated by the PM is smaller than the size of the tree based on the reduct generated by the RSM. So, the classification process when using the PM will be faster. Another comparison was made based on the coverage percentage which represents the ratio of classified objects (recognized by classifier) from the class to the number of all objects in the class. The coverage given by DTreeC was 90.2% for the RSM and it was 93.4% for the PM. These two comparisons have shown that the PM is better than the RSM when it was applied to the HSV dataset.

5.2. Two Levels of Comparisons

In this study, I established two levels of comparisons in order to test the proposed model results. The first-level comparison which based on the cardinality is the main comparison used in this study in order to decide which the best model is (between the rough set model and the proposed model). If both models give the same cardinality of reducts, the second-level comparisons (the accuracy, the coverage percentage and/or the number of nodes) could be considered.

Table 5 summarizes all comparisons between the RSM and the PM. The comparisons are extended to cover first-level and second-level comparisons. Comparisons were applied to each dataset as needed. In Table 5, the last column evaluates the PM based on one or more of the mentioned comparisons. If the PM passes the first-level comparison then we do not have a need to go for further comparisons (second-level comparison) as in Hepdata and Heart-Disease datasets. If the result of the first-level comparison (reduct's size) is the same for both models as in Dermatology and HSV datasets then second-level comparison is established. The main priority for the second-level comparisons is the accuracy of the model followed by the coverage and finally the number of nodes (note that we may follow different sequence of priorities as needed). For dermatology dataset, the PM pass-test since its accuracy is better than the RSM in both RBC and DTreeC. Note here that there is no need to go further in testing the other second-level comparisons since the PM is already passed-test. HSV dataset has the same accuracy for both RBC and DTreeC. So, we go further and test the coverage given by both techniques. The coverage given by RBC is the same for both models so we calculate the number of nodes given by the RSES for HSV dataset. The results showed that the number of nodes given when the reduct of the PM was employed is less than number of nodes given when the reduct of the RSM was employed. Consequently, the PM is passed-test for HSV dataset. The coverage given by DTreeC when it was applied to HSV dataset using the PM was better than that of the RSM. This is another proof that the PM is also passed-test for HSV dataset.

Table 5. First and second levels of comparisons; NA: not applicable to be used here (no effect).

DATASET	First Level	Second Level							Evaluation of the PM
	Reduct's size	The testing Techneques	Accuracy		Coverage		# of nodes		
			RSM	PM	RSM	PM	RSM	PM	
Hepdata	Less	NA							Pass
Dermatology	Same	RI	86.4	97.3	NA				Pass
		DTree	87.7	97.2					
Heart-Disease	Less	NA							Pass
HSV	Same	RI	NA				7	3	Pass
		DTree	NA	90.2	93.4		7	3	Pass

6. Conclusions

Generating the minimal reduct of a dataset which has a lot of missing values is a challenge. Missing data may exist because of the unavailability of data or because of security purposes. It has been noted that building data mining systems from incomplete data seems harder than that of complete data. In this article, a model was proposed and an algorithm was built to generate the minimal redut of incomplete medical datasets. The quality of the reduct generated is also researched. The model was tested against different measures such as cardinality, accuracy, number of nodes, and coverage percentage.

Comparisons between the results of the PM and the results of the RSM were made based on the first-level and second-level comparisons as needed.

The PM basically imputed the missing values of the original dataset then it splits the original dataset into two datasets: DS1 and DS2. The reducts were generated for DS1 and DS2. The PM has merged the reducts of DS1 and DS2 and then found their minimal reduct. The resulted dataset based on the minimal reduct was trained against RBC and DTreeC.

The PM grew the advantage over rough set reduct since it always gives the minimal reduct (limited to the incomplete datasets that can be vertically distributed such as the medical datasets). It also provided high accuracy model (generated from the reduct) which is the same or higher than that of rough set model subject to the given dataset. The PM produced better results than the model of rough set.

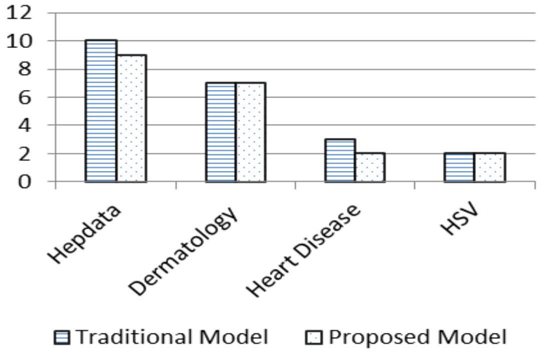

Figure 3. Cardinalities comparison.

Acknowledgment

This research was funded by the Arab Open University/Kuwait Branch under the Grant No. 84/2015. The author likes to thank many anonymous people for their efforts in improving the readability of this paper including the patience of my wife and kids.

References

[1] Agrawal A. and Srikant R., “Privacy Preserving Data Mining,” *in Proceedings of the ACM SIGMOD International Conference on Management of Data*, Dallas, pp. 439-450, 2000.

[2] Al-Shalabi L., Mahmod R., Abd Ghani A., and Saman Y., “A New Model for Extracting a Classifactory Knowledge from Large Datasets Using Rough Set Approach,” *in Proceedings of World Engineering Congress*, Kuala Lumpur, pp. 105-108, 1999.

[3] Al Shalabi L., Shaaban Z., and Kasasbeh B., “Data mining: A Preprocessing Engine,” *Journal of Computer Science*, vol. 2, no. 9, pp. 735-739, 2006.

[4] Bazan J., Szczuka M., Wojna A., and Wojnarski M., “On Evolution of Rough Set Exploration System,” *in Proceedings of International Conference on Rough Sets and Current Trends in Computing*, Berlin, pp. 592-601, 2004.

[5] Berthold M. and Huber K., “Missing Values and Learning of Fuzzy Rules,” *International Journal Uncertainty, Fuzziness Knowledge-Based Systems*, vol. 6, no. 2, pp. 171-178, 1998.

[6] Chen D., Zhang L., Zhao S., Hu Q., and Zhu P., “A Novel Algorithm for Finding Reducts with Fuzzy Rough Sets,” *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 2, pp. 385-389, 2012.

[7] Dai J., Wang W., Tian H., and Liu L., “Attribute Selection based on a New Conditional Entropy for Incomplete Decision Information Systems,” *Knowledge-Based Systems*, vol. 39, pp. 207-213, 2013.

[8] Dempster A., Larid N., and Rubin D., “Maximum Likelihood from Imcomplete Data via the Em Algorithm (with Discussion),” *Journal of Royal Statistical Society*, vol. 39, pp. 1-38, 1977.

[9] Inuiguchi M., Yoshioka Y., and Kusunoki Y., “Variable-precision Dominance-based Rough Set Approach and Attribute Reduction,” *International Journal of Approximate Reasoning*, vol. 50, no. 8, pp. 1199-1214, 2009.

[10] Jacob S. and Raju G., “Software Defect Prediction in Large Space Systems through Hybrid Feature Selection and Classification,” *The International Arab Journal of Information Technology*, vol. 14, no. 2, pp. 208-214, 2017.

[11] Jia X., Liao W., Tang Z., and Shang L., “Minimum Cost Attribute Reduction in Decision-theoretic Rough Set Models,” *Information Sciences*, vol. 219, pp. 151-167, 2013.

[12] Meng Z. and Shi Z., “Extended Rough Set-Based Attribute Reduction in Inconsistent Incomplete Decision Information Systems,” *Information Sciences*, vol. 204, pp. 44-69, 2012.

[13] Michikazu N. and Weiming K., “Review of the Methods for handling Missing Data in Longitudinal Data Analysis,” *International Journal of Math. Analysis*, vol. 5, no. 1, pp. 1-13, 2011.

[14] Parthalain N., Shen Q., and Jensen R., “A distance Measure Approach to Exploring the Rough Set Boundary Region for Attribute Reduction,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 3, pp. 305-317, 2010.

[15] Pawlak Z. and Skowron A., “Rudiments of Rough Sets,” *Information Sciences*, vol. 177, no. 1, pp. 3-27, 2007.

[16] Quinlan J., “Unknown Attribute Values in Induction,” *in Proceedings of the 6th International Workshop on Machine Learning*, Ithaca, pp. 164-168, 1989.

[17] Rough Set Exploration System (RSES),

http://www.mimuw.edu.pl/~szczuka/rses/start.htm, Last Visited, 2015.

[18] Sansom C., "Up in a Cloud?," *Nature Biotechnology*, vol. 28, no. 1, pp. 13-15, 2010.

[19] Taylor R., "An Overview of the Hadoop/Mapreduce/Hbase Framework and Its Current Applications in Bioinformatics," *in Proceedings of the 11th Annual Bioinformatics Open Source Conference*, Boston, 2010.

[20] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/, Last Visited, 2015.

[21] Ye D., Chen Z., and Ma S., "A Novel and Better Fitness Evaluation for Rough Set based Minimum Attribute Reduction Problem," *Information Sciences*, vol. 222, pp. 413-423, 2013.

[22] Zhang S., "Shell-Neighbor Method and its Application in Missing Data Imputation," *Applied Intelligence*, vol. 35, pp. 123-133, 2011.

[23] Zhong M. and Sharma S., "Development of Improved Models for Imputation Missing Traffic Counts," *The Open Transportation Journal*, vol. 3, pp. 35-48, 2009.

Luai Al Shalabi was born in Jordan in 1971. He received the B.S. in computer science from Yarmouk University, Jordan, in 1992, M.S. degrees in image interpretations from Universiti Sains Malaysia, Malaysia, in 1996, and the Ph.D. degree in data mining from University Putra Malaysia, Malaysia, in 2000. He is working in the Information Technology Department at Arab Open University in Kuwait. His research interests include data mining, knowledge discovery, and machine learning.