Study of Assigning Imperfect Attribute Values for Classifier

Thi Thi Soe¹ and Zarni Sann²

¹Faculty of Computer Science,
²Faculty of Computer Systems and Technologies
University of Computer Studies (Mandalay)
Myanmar

ABSTRACT

One of the interesting and important fields of research in data mining is classification on imperfect data. Unknown value can appear in real-world data sets at the stage of data collection. These facts lead to the imperfection of the decision system. The attention of this paper is to study the methodologies in making a decision point on such incompleteness of data sets. We found a well-known rough set (RS) based classification scheme: Learning from Examples based on Rough Sets (LERS) that could be treated missing data, numeric data, and inconsistent data set. In this study, we utilize two interpreted meaning of imperfect values: lost values and attribute-concept values. The classification system is illustrated using a case study of iris dataset from the UCI repository. The system is intended to present a comprehensive view of assigning on imperfect attributes value that generates better result among lost and attribute-concept values.

Keywords: Data mining, Imperfect data, lost value, Attribute–concept value, Rough set-based Classifier.

1. INTRODUCTION

Mathematical rough set theory was first appeared in the pioneering work in [1]. The idea of rough set was successfully implemented and used in many applications such as classification, pattern recognition, learning algorithms. Classical rough set can only deal with complete information systems in which the data available. However unknown values can attach in real world data at the stage of data gathering. Imprecise and imperfection of data table may affect the performance of pattern analysis and classification system. Hence concentration of many researchers turned to develop algorithms for attenuate such type of problems. Large bodies of attempts extending the rough set have been demonstrated in knowledge acquisition by discovering rules set when presented a datasets with missing data. Among them, the most successful one is system LERS introduced in [2]. LERS applied the algorithms: Learning from Examples Module version1 and version2, LEM1 and LEM2 [2, 3, 4]. Global and local coverings were used in LEM1 and LEM2 respectively. For decision rule generation, LEM2 algorithm was successfully implemented and used in [5-6]. Other approaches using LEM2 algorithm were described in [7-10].

Because of a major inconvenience in LEM2 on processing numerical attribute values, an approach is extended on it [11]. This extended version is called modifications of LEM2 or simply MLEM2. Later several researches were further presented to the fields of analyzing numerical data sets and missing attribute values [12-14]. LERS could also be manipulated inconsistent data set in which two or more cases with all of same attribute values have different decision value. Until current time the researcher continuously analyzing the data mining fields by utilizing system LERS. For example, in resent research [15] comparison results of mining incomplete and inconsistent data were presented in terms of an error rate. Experiments were conducted on 204 data sets and used rule induction of MLEM2 algorithm. They concluded that incompletenss was worse than inconsistency for data mining. Missing type lose values were better than “do not care conditions”. Hence RS-based classifier is leading to a fruitful research and applications in many places. For comparison purpose an RSFit approach was introduced in [16] to assign missing attribute values from rough sets perspective. They demonstrated the RSFit approach by an artificial car data set in [17], UCI data sets and a Geriatric care data set with randomly selected missing attribute values. The accuracy of the prediction was compared to the “closest fit” approach proposed by Grzymala-Busse. They point out that the RSFit approach significantly reduced the computation time and achieved comparable accuracy result.
Generally, the methods for incomplete data classification based on attributes require extra computing: imputation or classifier updating. Imputation is a procedure that consists of assigning the missing values with predicted ones based on available information in the data set. These methods lead to be computational burden. Reference [18] distinguished the missing data into three types: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). There have been proposed other approaches on how to manipulate missing data in the literature. Reference [19] proposed an approach, using the entropy to estimate the missing values. Experimental results showed that this approach could be able to outperform other imputation techniques, as mean and mode. In [20] a partial imputation technique was developed. It involves the imputation of missing data using complete objects in a small neighborhood of the incomplete ones. In [21] analyzed the efficiency for missing data treatment using classification-based mining algorithm C4.5 and cluster-based mining algorithm K-means. In these experiments, missing values were artificially imputed, in different rates and attributes, into the numerical data sets. K-means imputation method provided good results in most cases of their experiments.

Support vector machines (SVMs) [22] are learning models frequently used for classification. In [23], the imputation technique using SVM was studied and compared it with the traditional imputation techniques mean, median, the mean of the two closest neighbor values and the value of the nearest neighbor techniques. Their results showed that the SVM technique obtained the highest precision with regards to the others methods. Reference [24] examined the impact of performing missing data imputation with five imputation methods: mean, the Hot deck method, Naïve Bayes, multiple imputation and framework method using Naïve Bayes and Hot deck. Experimental analysis was tested with six modern classifiers (RIPPER, C4.5, KNN, SVM, and Naïve Bayes) on 15 data sets to investigate the effect of imputation on the classification errors. Performed analysis result shows that imputation methods are beneficial except mean imputation for the classification of objects with missing attributes.

According to a brief overview of incomplete data handling techniques discussed in [25], the missing data imputation approaches based on machine learning, artificial neural network algorithms, K-nearest neighbor algorithm and Self-Organizing Maps (SOM) are more frequently used. On the other hand three techniques: Hot deck imputation, mean substitution, and Regression substitution are rarely used due to poor classification performance. Among previous classification methodologies when analyzing datasets with missing data, our preferences are on the series of Grzymala-Busse’s data mining tasks [11, 12, 14, 26-29]. Hence the goal pursued by this paper is to create a Java code program for LERS–classification framework. The goal of this paper is also to give a comprehensive view related to the meaning of imperfect attribute values: lost and attribute-concept values by providing implementation results.

2. METHOD AND MATERIAL

2.1 Incomplete Information systems

Dealing with the rough set concept, an information system, I, is defined as a pair \( I = (U, A) \), where \( U \) is a non-empty finite set of objects (cases) called the universe and \( A \) is a non-empty finite set of attributes such that \( a : U \rightarrow V_a \) for every \( a \in A \). The set \( V_a \) is called the value set of \( a \). A decision system is special form of information system such that \( I = (U, A \cup \{d\}) \), where \( d \) is the decision attribute. The elements of \( A \) are called conditional attributes or simply conditions [30]. Decision tables can be utilized as the training data sets for data mining tasks. It is possible in an information system some positions of attribute values are empty or absent. This called an incomplete information system [31, 32]. Also refers to imperfect decision table as described in Table 1.

<table>
<thead>
<tr>
<th>Case</th>
<th>Attributes</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Att.1</td>
<td>Att.2</td>
</tr>
<tr>
<td>1</td>
<td>a11</td>
<td>a13</td>
</tr>
<tr>
<td>2</td>
<td>a21</td>
<td>a23</td>
</tr>
<tr>
<td>3</td>
<td>a31</td>
<td>a32</td>
</tr>
</tbody>
</table>

2.2 Meaning of Missing values

Absent positions can be assigned with a value based on the available information of data set. Grzymala-Busse and other researchers [33, 26, 27, 28, 29] have been devoted the meaning of missing attribute values in incomplete decision table. As
described in Table 2 they defined the definitions of missing types along with the relevant symbols and distinguished how to assign a value on each symbol for successive RS-based data mining tasks.

<table>
<thead>
<tr>
<th>Definition</th>
<th>Symbol</th>
<th>Assign type</th>
<th>Assume</th>
</tr>
</thead>
<tbody>
<tr>
<td>lost value</td>
<td>?</td>
<td>not assigned by any existent value from the attribute domain</td>
<td>erase</td>
</tr>
<tr>
<td>“Do not care” condition</td>
<td>*</td>
<td>assigned by any possible value from the attribute domain</td>
<td>unnecessary</td>
</tr>
<tr>
<td>Attribute-concept value</td>
<td>_</td>
<td>assigned by any possible value from the attribute domain bounded to the decision value</td>
<td>has a concept</td>
</tr>
</tbody>
</table>

2.3 Classifier–LERS

Rule induction is one of the important steps in machine learning classification. The basic rule induction algorithm for classifier-LERS is MLEM2. MLEM2 could be able to discover rules from numerical and symbolic attribute of imperfect decision table [11].

2.3.1 Transforming numeric-values

To induce rules from decision table, the first step is discretization of numerical attributes. Doing so required the transformation processes listed below [12].

- sorts all numeric values excluding missing attribute values
- computes cut-points as averages for any two consecutive values of the sorted list
- creates two blocks: all cases with numeric value <\( C \), and all cases with numeric value >\( C \)

2.3.2 Central blocks

Central blocks: the attribute-value pair block is the direct basic for RS-based classifier. The attribute-value pair block \([a, v]\) is defined as the set \(\{x \in U | a(x) = v\}\) [34]. For imperfect decision tables, the definition of the block is modified according to the missing types described in Table 2. The block is used to derive the characteristics relation, characteristics set, and lower and upper approximations. We found that imperfect data sets are usually analyzed using three classes of approximations such as singleton, subset and concept [12, 6]. In our implementation, we used the concept definition of \(\hat{A}\)-lower and \(\hat{A}\)-upper approximation for discovering certain and possible rules set with equations, \(\hat{A}_{\text{concept}}(X) = U\{K_a(x)|x \in X, K_a(x) \subseteq X\}\) and \(\hat{A}_{\text{concept}}(X) = U\{K_a(x)|x \in X, K_a(x) \cap X \neq \emptyset\} = U\{K_a(x)|x \in X\}\), where \(X\) is a subset of \(U\), and \(K_a(x)\), the characteristic set is the intersection of the sets \(K(x, a)\), for all \(a \in A\). For classify testing data and unknown cases, classifier-LERS utilizes the set of rules induced by the algorithm [14, 12].

3. IMPLEMENTATION

In order to illustrate the study area we create Java code program on iris plant dataset from UCI database [35]. The data set contains 150 cases in which 50 in each of three classes. The process flow of classification system on imperfect data is followed by LERS- classifier [12, 11, 14]. In this experiment, the data set is divided into 120 cases for training and 30 cases for testing according to the holdout method. Training data sets are built by randomly placing the lost values at 20 and 40 positions for representing the imperfection of data. Attribute-concept values are also placed at the same positions. The blocks derive as a vital engine at the initial state of computing characteristic set until generating the certain and possible rules based on the concept lower and upper approximation. Rules generated by MLEM2 algorithm for 20 numbers of ‘attribute-concept’ missing values are presented to the user as shown in Figure 1.

![Figure 1](image-url)
These rules are employed by LERS classifier to classify thirty cases of testing data. The quality of classifier is measured using the following equation: Testing data and the result of calculated accuracy is given in Figure 2.

\[
\text{classification accuracy} = \frac{\text{number of correctly classified cases}}{\text{total number of testing cases}} \times 100
\]  

(1)

![Figure 2: Classifier accuracy for 20 numbers of ‘attribute-concept’ missing values](image1)

And then, the system tested with 40 numbers of attribute-concept values, 20 and 40 numbers of lost values as in similar ways. The performance of the classifier is reported according to the number of imperfect data for each missing types. Figure 3 shows the classifier accuracy with the bar graph representation for comparison purpose.

![Figure 3: Comparisons of classifier accuracy for different missing types](image2)

The experiment also reports the generated number of rules and conditions tested on each of the missing types. Results are tabulated in Table 3.

<table>
<thead>
<tr>
<th>Number of missing value</th>
<th>Lost</th>
<th>Attribute-concept</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>number of rules</td>
<td>number of conditions</td>
</tr>
<tr>
<td>20</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>40</td>
<td>10</td>
<td>24</td>
</tr>
</tbody>
</table>
4. CONCLUSION

This paper studied the methods of assigning on imperfect and uncertain data for classification task. For treating absent and uncertain data, rough set-based classifier provides as a great mechanism. Concerning with the study area in this paper we implemented the classifier which is based on the operations of LERS-system. This paper also analyzed the classifier accuracy by assigning values on imperfect data deals with the description of lost and attribute-concept values. Classifier utilizing lost values gained the better accuracy than the attribute-concept values both for different numbers of imperfect data: 20 and 40. The resulted accuracy rates also meet the solution of classification on incomplete data by previous works. Obtaining knowledge on studied domain classification model will be constructed combining rough set and other approaches in the future.

REFERENCES


