Experiments on Incomplete Data Sets Using Modifications to Characteristic Relation

By

Sumiah A. Alalwani

Submitted to the Department of Electrical Engineering and Computer Science and the
Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Master of Science

Dr. Jerzy W. Grzymala-Busse, Chairperson

Committee members

Dr. Prasad Kulkarni

Dr. Bo Luo

Date defended:

The Thesis Committee for Sumiah A. Alalwani certifies
that this is the approved version of the following thesis :

Experiments on Incomplete Data Sets Using Modifications to Characteristic Relation

Dr. Jerzy W. Grzymala-Busse, Chairperson

Date approved:

Abstract

Rough set theory is a useful approach for decision rule induction which is applied to large life data sets. Lower and upper approximations of concept values are used to induce rules for incomplete data sets. In our research we will study validity of modifications suggested to characteristic relation. We discuss the implementation of modifications to characteristic relation, and the local definability of each modified set. We show that all suggested modification sets are not locally definable except for maximal consistent blocks that are restricted to data set with "do not care" conditions. A comparative analysis was conducted for characteristic sets and modifications in terms of cardinality of lower and upper approximations of each concept and decision rules induced by each modification. In this research, experiments were conducted on four incomplete data sets with lost and do not care conditions. LEM2 algorithm was implemented to induce certain and possible rules from the incomplete data set. To measure the classification average error rate for induced rules, ten-fold cross validation was implemented. Our results show that there is no significant difference between the qualities of rule induced from each modification.

Acknowledgements

I would like to express my deepest gratitude to my advisor Dr. Jerzy Grzymala-Busse for his persistent guidance and continuous support and encouragement on my study and thesis. His enthusiasm of teaching and research inspired me, and will have long lasting effects on me. I also would like to thank Dr. Bo Luo and Dr. Prasad Kulkarni for consenting to be member of my committee. Above ground, I am indebted and forever grateful to my parents, whose value to me only grows with age. Finally, I thank my brothers and sisters for always encouraging and supporting me.

Contents

List of Tables

Chapter 1

Introduction

Rough set theory is a useful approach for decision rule induction which is applied to large life data sets. Rough set approach is used to handle incomplete data sets, where there are two interpretations of missing attribute values: lost values, denoted by "?" which means that the original attribute value was known, however due to various reasons it was erased or never obtained. "Do not care" values are denoted by "*". Such value is irrelevant and can be replaced by any value of attribute domain, since its values does not affect the final outcome. Rough Set theory elementary sets are extended to deal with incomplete data sets where characteristic set and characteristic relation are proposed by Jerzy Grzymala-Busse to deal with incomplete information system with both lost and "do not care" conditions. The characteristic set $K_B(x)$ may be interpreted as the set of all cases that are indistinguishable from object x and for all attributes $a \in B$ where the attribute value pairs are based on the interpretation of missing attribute values. The characteristics relation R(B) is reflexive but not symmetric or transitive. Decision rule induction is the process by which rules are induced from the decision tables. It involves extraction of high level information from low level data and it is the most fundamental data mining technique. Rule induction algorithm LEM2 is used with lower and upper approximations of concept values to induce rules for incomplete data sets. In this thesis, we will study modifications to the definition of characteristic relation that were suggested due to two unreasonable situations, first the characteristic relation is classifying two objects that do not have any known equivalent attribute values to be in the same class or two objects that have a lot of known equal attribute values, but are members of different classes. Another approach for rule induction on incomplete data sets is the concept of maximal consistent block. It is defined as that maximal collection of objects in which all objects are similar, and they are indiscernible based on the attribute values available. The binary relation of maximal consistent blocks is symmetric and reflexive.

The objective of this thesis is to study validity of modifications suggested to characteristic relation and the local definability of each modified set, as well as the impact of each modification on rule induction. A comparative analysis was conducted for characteristic sets and modifications in terms of cardinalities of lower and upper approximations of each concept and decision rules induced by each modification. Also, we measured the classification average error rate for induced rules by implementing ten-fold cross validation.

Section 2 covers background knowledge on rough set theory and information systems. Section 3 covers modifications to the characteristic relation, maximal consistent blocks and the local definability of each modified relation characteristic sets. Section 4 covers data sets used in conducting experiments and results of the experiments. Section 5 covers the conclusion derived from the experiments.

Chapter 2

Background

This chapter will provide an overview of complete and incomplete information systems, fundamental concepts of rough set theory and characteristic relation.

2.1 Information Systems and Decision Table

Decision table [13] is used as a way to represent knowledge and data in data mining, in which each input data set is represented as a decision table. Each row of the decision table will represent an object, and each column corresponds to a variable called an attribute, where attribute values will provide information about the object, and attribute decision value will classify the object to a concept. Conventionally, information system is the duple IS = <U, A> where U is non-empty finite set of objects called the universe, and A = C $\cup$ D is a non-empty finite set of attributes, C is the set of attributes and D is the set consisting of a decision attribute. For every a $\in$ A and x $\in$ U, we have the a(x) $\in$ Va where Va is called the domain of the attribute a. V = $\bigcup_{a\in A}$ V_a is the value set of all attributes. Let a $\in$ A, x $\in$ U, v $\in$ V_a, then the pair (a,v) is called attribute-value pair, then the block of an attribute-value pair is denoted by [(a,v)] for a complete data set is [(a,v)]= $\{$ x | x $\in$ U, a(x) = v $\}$.

2.2 Rough Set Theory

Rough Set Theory, introduced by Z.Pawlak is a mathematical tool used for analyzing the uncertainty and vagueness of a data model. It is proven to be useful in Artificial Intelligence, Data Mining, Machine Learning, Pattern Recognition, Knowledge Acquisition, etc. The main goal of rough set analysis is the induction of approximation of concepts that can be used for feature selection, feature extraction, data reduction, decision rule generation, and pattern extraction.

Information system is complete if there is no missing attribute values and consistent if there do not exist any objects with the same values to all attribute conditions, but their decision attribute has different values. In rough set theory complete and consistent information system is categorized by the indiscernibility relation [6], which is denoted by IND(B), where $B \subseteq A$, A is the set of all attributes. $IND(B) = \{ (x, y) \mid (x, y) \in U \times U \text{ and } \forall a \in B \Rightarrow a(x) = a(y) \}$,

where a(x) denote the value of the attribute for the object x. Indescribability relation describes the indescribability of an object, which means that any two objects are indistinguishable from each other, therefore the indescribability relation on a complete data set is an equivalence relation. The equivalence relation is reflexive, symmetric and transitive. The equivalence classes of B-Indescribability relation are called elementary sets of B and denoted by $[x]_B$. Any union of B-elementary sets will be called a B-definable set.

For complete data set, the lower and upper approximations [13] will be defined by using elementary sets of R(B). Let X be a concept, $B \subseteq A$, $x \in U$ and R(B) is the equivalence relation for complete data set which its elementary sets are $[x]_B$, the B-lower and B-upper approximations of X are defined as follows:

$\underline{B}X = \cup\{[x]_B \mid x \in X, [x]_B \subseteq X\}$, $\overline{B}X = \cup\{[x]_B \mid x \in X, [x]_B \cap X \neq \emptyset \} = \cup\{[x]_B \mid x \in X\}$.

2.3 Incomplete Information Systems

In the real world, there are many incomplete data sets with missing attribute values due to various reasons. There are two known interpretations of missing attribute values:

- Lost values, denoted by ?

  The original attribute value was known, however due to various reasons it was erased and never obtained. If for an attribute a, there exists a object x where a(x) = ?, then x is not included in any [(a, v)] blocks for all specified values v of attribute a.

- "Do not care" values, denoted by *

  Such attribute value can be replaced by any value of attribute domain, since it does not affect the final outcome. If for an attribute a, there exists a object x where a(x) = *, then x is included in all [(a, v)] blocks for all specified values v of attribute a.

object	Temperature	Headache	Nausea	Flu
1	High	?	No	Yes
2	Very-high	Yes	Yes	Yes
3	?	No	No	No
4	High	Yes	Yes	Yes
5	High	?	Yes	No
6	Normal	yes	No	No
7	Normal	No	Yes	No
8	*	Yes	*	Yes

Table 2.1: An example of incomplete decision table

The attribute-value blocks for Table 2.1 data set when B = A are shown in Table 2.2.

(a,v)	[(a,v)]
(Temperature, High)	{1, 4, 5, 8}
(Temperature, Very-high)	{2, 8}
(Temperature, normal)	{6, 7, 8}
(Headache, yes)	{2, 4, 6, 8}
(Headache, no)	{3, 7}
(Nausea, no)	{1, 3, 6, 8}
(Nausea, yes)	{2, 4, 5, 7, 8}

Table 2.2: Attribute-value blocks for Table2.1

2.4 Characteristic Sets and Characteristic Relation

Rough Set theory elementary sets are extended [3][9] to deal with incomplete data sets, characteristic set and characteristic relation are introduced by Jerzy Grzymala-Busse to deal with incomplete information system with both lost and "do not care" conditions. For object x $\in$ U and B $\subseteq$ A, the characteristic set K_B(x) of B is defined as the intersection of the sets K(x, a). For all a $\in$ B, where the set K(x, a) is defined in the following way:

(1) If a(x) is specified, then K(x, a) is the block [(a, a(x))] of attribute a and its value a(x).

(2) If a(x) =? or a(x) = *, then the set K(x, a) = U, where U is the set of all objects.

The characteristic sets for Table 2.1 incomplete data set where B = A are shown in Table 2.3.

K_B(x)
K_B(1) = {1, 8}
K_B(2) = {2, 8}
K_B(3) = {3}
K_B(4) = {4, 8}
K_B(5) = {4, 5, 8}
K_B(6) = {6, 8}
K_B(7) = {7}
K_B(8) = {2, 4, 6, 8}

Table 2.3: Characteristic sets for Table 2.1.

For incomplete data set, the lower and upper approximations [3] will be defined by using characteristics sets instead of elementary sets. For incomplete data sets there are three definitions of lower and upper approximation : singleton, subset and concept. Let X be a concept, B $\subseteq$ A, x $\in$ U, and R(B) be a characteristic relation for incomplete data set with characteristic set K(x), then the singleton B-lower and B-upper approximations of X are defined as follows:

$\underline{B}$X = {x $\in$ U | K_B(x) $\subseteq$ X}, $\overline{B}$X = {x $\in$ U | K_B(x) $\cap$ X $\neq \emptyset$}

For incomplete data set presented in Table 1, the singleton A-lower and A-upper approximations of the two concepts {1, 2, 4, 8} and {3, 5, 6, 7} are:

$\underline{A}${1, 2, 4, 8}={1, 2, 4}, $\underline{A}${3, 5, 6, 7}={3,7}.

$\overline{A}${1, 2, 4, 8}={1, 2, 4, 5, 6, 8}, $\overline{A}${3, 5, 6, 7}={3, 5, 6, 7, 8}.

A subset B-lower and B-upper approximations of X are defined as follows:

$\underline{B}$X = $\cup${K_B(x) | x $\in$ U, K_B(x) $\subseteq$ X}, $\overline{B}$X = $\cap${K_B(x) | x $\in$ U, K_B(x) $\cap$ X $\neq \emptyset$}.

For incomplete data set presented in Table 1, the subsets A-lower and A-upper approximations of the two concepts are:

$\underline{A}${1, 2, 4, 8}={1, 2, 4, 8}, $\underline{A}${3, 5, 6, 7}={3,7}.

$\overline{A}${1, 2, 4, 8}={1, 2, 4, 5, 6, 8}, $\overline{A}${3, 5, 6, 7}={2, 3, 4, 5, 6, 7, 8}.

A concept B-lower and B-upper approximations of X are defined as follows:

$\underline{B}$X = $\cup${K_B(x) | x $\in$ X, K_B(x) $\subseteq$ X},

$\overline{B}$X = $\cup${K_B(x) | x $\in$ X, K_B(x) $\cap$ X $\neq \emptyset$}=$\cup${K_B(x) | x $\in$ X}.

For incomplete data set presented in Table 2.1, the concept A-lower and A-upper approximations of the two concepts are:

$\underline{A}${1, 2, 4, 8}={1, 2, 4, 8}, $\underline{A}${3, 5, 6, 7}={3, 7}.

$\overline{A}${1, 2, 4, 8}={1, 2, 4, 6, 8}, $\overline{A}${3, 5, 6, 7}={3, 4, 5, 6, 7, 8}.

By definition, subset and concept lower and upper approximations are globally definable; therefore they may be applied in data mining.

The characteristics relation R(B) is a relation on a set of all objects U defined for x, y $\in$ U as follows: (x, y) $\in$ R(B) if and only if y $\in$ K_B(x)

Characteristic relation for Table 2.1 data set, where B = A defined as

R(A)={(1, 1), (1, 8), (2, 2), (2, 8), (3, 3), (4, 4), (4, 8), (5, 4), (5, 5), (5, 8), (6, 6), (6, 8), (7, 7), (8, 2), (8, 4), (8, 6), (8, 8)}.

The characteristics relation R(B) is reflexive but not symmetric or transitive. The characteristic relation is known if we know all the characteristic sets x $\in$ U. When all missing attribute values are lost values, the characteristic relation R(B) is reflexive and transitive and the relation will be called a similarity relation. When all missing attribute values are "do not care" conditions, then the B-characteristic relation R(B) is reflexive and symmetric, and the relation will be called a tolerance relation.

2.5 Rule Induction

Decision rules classify data in the decision table to different concepts. A rule is represented in the LERS [4][13] format as:

Number$_1$, Number$_2$

(attribute$_1$, $value_1$)&(attribute$_2$, $value_2$)&....&(attribute$_n$, $value_n$)$\rightarrow$(decision, value)

The left hand side of the rule represents the attribute-value pairs and the right hand side represents the concept. A object x $\in$ U is covered by a rule *r* if and only if every condition of *r* is satisfied by the corresponding attribute *a* with value *v* for object *x*. The numbers Number$_1$ and Number$_2$ are based on LERS classification System in which they represent Strength and Specificity. Strength is defined as the total number of training objects that are correctly classified by the rule. If all the objects that are covered by rule*r* are correctly classified, then the rule *r* is consistent with the dataset. Specificity is the total number of conditions or attribute value pairs in the rule *r*. Decision rule induction is the process by which rules are induced from the decision tables. It involves extraction of high level information from low level data and is the most fundamental data mining technique. Examples for rule induction algorithms are LEM1 and LEM2 .

2.5.1 The LEM2 Algorithm

The LEM2 (Learn from Examples Module, version 2) algorithm is a module of the LERS learning system. LEM2 computes local covering for every concept of the training data set which will be converted to a rule set [4][13]. For Incomplete data set, LEM2 will compute the local covering of the approximations of every concept.

Let B be nonempty lower or upper approximation of a concept. T is a set of attribute-value pairs where each (a, v) = t, a block of t is denoted by [t], which is a set of all objects $x \in U$ that attribute a have value v. T is a minimal complex of B only if

$$[T] = \bigcap_{t \in T} [t] \subseteq B$$

and there does not exist a subset T' of T such that T' $\subseteq$B. Let $\boldsymbol{T}$ be a nonempty collection attribute-vlaue pairs T, then $\boldsymbol{T}$ is a local covering if and only if it satisfies the following conditions:

(1) each member T of $\boldsymbol{T}$ is a minimal complex of B,

(2) $\bigcup_{T \in \boldsymbol{T}}$=B,

$\boldsymbol{T}$ is minimal, such that if any $T \in \boldsymbol{T}$ is removed, then second condition is not satisfied.

Consider an incomplete decision Table 2.1, a local covering of the characteristic set upper approximation of concept [(Flu, yes)], $\overline{A}\{1, 2, 4, 8\} = \{1, 2, 4, 6, 8\}$ is

$\boldsymbol{T}$ = {{(Headache, yes)},{(Temperature, high), (Nausea, no)}} which corresponds to the following possible rules :

3,1

(Headache, yes) → (Flu, yes)

1,2

(Temperature, high) & (Nausea, no) → (Flu, yes)

where the first rule covers objects {2, 4, 6, 8} and the second rule covers objects {1, 8}. These two rules are consistent with the data set, and all of the objects of upper approximation of concept[(Flu, yes)] are covered. Also, all rules induced are minimal and there is no redundant rule. A detailed description of LEM2 algorithm is given in Appendix A.

2.5.2 Global and Local Definability

For incomplete data sets, a set X $\subseteq$ U is B-globally definable if it is a $\bigcup_{K_{\mathrm{B}}(x)}$, x $\in$ U. If a set is A-globally definable, then it will be called globally definable. Local definability is based on attribute-values pairs granules. A set T of attribute-value pairs, where all attributes are distinct and belong to a set B, a subset of the set A, will be called a B-complex. A block of B-complex T, denoted by [T], is defined as the set $\cap$ $\{$ [t] | t $\in$ T$\}$.

Let B be a subset of U. Set B depends on a set T of attribute-value pairs where each t = (a, v) if and only if [T] is nonempty and [T] $\subseteq$ B.

In incomplete decision table, Let B $\subseteq$ A where $\bigcup_{T \in \boldsymbol{T}}$ from some B-complexes will be called a B-locally definable set. A-locally definable sets will be called locally definable. An approximation of a concept should be locally definable since decision rules are expressed in the form of attribute-value pairs as noted above. In the listed conditions of local covering, condition $\bigcup_{T \in \boldsymbol{T}}$ = B means that the set [(decision, value)] must be locally definable. Since concepts approximations are used to induce rules of the incomplete data set, characteristic sets that are used to construct approximation must be at least locally definable. Any set X that is B-globally definable is B-locally definable as well, however not every B-locally definable is B-globally definable. For decision tables in which all missing attribute values are lost, local definability is reduced to global definability.

Chapter 3

Modifications to Characteristic Relation

This chapter will provide an overview of all suggested modifications to the definition of characteristic relation, also we will study the local definability of each modified relation characteristic sets. A detailed description of the implementation of the modifications to characteristic relation is given in Appendix B.

3.1 Motivation for the modification

The modifications to the definition of characteristic relation were suggested due to two unreasonable situations, the characteristic relation is classifying two objects that do not have any known equivalent attribute values to be in the same class or two objects that have a lot of known equal attribute values to be in different classes. For example, for an incomplete decision Table 2.1, we see that objects 1 and 8 have no known equal attribute values, however object 8 belongs to $R_A(1)$. Also for objects 4 and 5 where they have a lot of known equal attribute values, but object 5 is not included in $R_A(4)$. Let S be an incomplete information system in which B $\subseteq$ A, x $\in$ U, For the next few definitions we will use the following notations:

$$P_B(x) = \{a \mid a \in B, a(x) \neq *\}, Q_B(x) = \{a \mid a \in B, a(x) \neq ?\},$$

$$C = Q_B(x) \cap QB(y) \neq \emptyset, D = P_B(x) \cap Q_B(y) \cap P_B(x) \cap Q_B(y) \neq \emptyset,$$

$$I_U \text{ is the identity relation on U} = \{(x,x) \mid x \in U\}.$$

3.2 R’ Characteristic Relation

In [11], the author proposed a characteristic relation defined by R’ to deal with the two unreasonable situations mentioned above. R’ characteristic relation is defined as follow:

$R'(B) = \{ (x, y) \in U \mid (\forall a \in C, a(x)=a(y) \vee a(x) = * \vee a(y) = *) \wedge (\forall a \in D, a(x) = a(y)) \} \cup I_U$.

R’ characteristic sets for Table 2.1 incomplete data set where B = A , are presented in Table 3.1.

$R'_A(x)$
$R'_A(1) = \{1, 3\}$
$R'_A(2) = \{2, 8\}$
$R'_A(3) = \{1, 3\}$
$R'_A(4) = \{4, 5, 8\}$
$R'_A(5) = \{4, 5, 8\}$
$R'_A(6) = \{6, 8\}$
$R'_A(7) = \{7\}$
$R'_A(8) = \{2, 4, 6, 8\}$

Table 3.1: R’ characteristic sets for Table 2.1 incomplete data set

R’ Characteristic relation for Table 2.1 data set, where B = A is defined as :

R’(A) = {(1, 1), (1, 3), (2, 2), (2, 8), (3, 1), (3, 3), (4, 4), (4, 5), (4, 8), (5, 4), (5, 5), (5, 8), (6, 6), (6, 8), (7, 7), (8, 2), (8, 4), (8, 6), (8, 8)}.

R’ is reflexive and symmetric while it is not necessary transitive. The existence of Identity Relation will ensure the reflexive property of R'_A.

A concept B-lower and B-upper approximations of X are defined as follows:

$\underline{A}X = \cup\{R'_A(x) \mid x \in X, R'A(x) \subseteq X\}$,

$\overline{A}X = \cup\{R'_A(x) \mid x \in X, R'_A(x) \cap X \neq \emptyset \} = \cup\{R'_A(x) \mid x \in X\}$.

For incomplete data set presented in Table 2.1, the concept A-lower and A-upper approximations of the two concepts are:

$\underline{A}\{1,2,4,8\} = \{2, 8\}$, $\underline{A}\{3,5,6,7\} = \{7\}$,

$\overline{A}\{1,2,4,8\} = \{1, 3, 2, 8, 4, 5, 6\}$, $\overline{A}\{3,5,6,7\} = \{1, 3, 4, 5, 6, 8, 7\}$.

3.3 R^1, R^2 and R^3 Characteristic Relations

In [12], characteristic relations R^1, R^2 and R^3 are proposed for solving different unreasonable situations in $R_A(x)$ where R^1, will discard objects that do not have any known attribute values to be classified in the same class. R^2 will include objects that have a lot of known equivalent attribute, but were discarded in $R_A(x)$ due to the existence of lost conditions to be in the same class. R^3 will combine the advantages of R^1 and R^2.

R^1characteristic relation is defined as follows:

$R^1(B) =\{ (x,y) \in UxU \mid (\forall\, a \in QB(x), a(y) \neq ?) \wedge (\forall\, a \in D, a(x) = a(y))\}$

R^1 characteristic sets for Table 2.1 incomplete data set, where B = A , are presented in Table 3.2.

$R^1{}_A(x)$
$R^1{}_A(1) = \{1\}$
$R^1{}_A(2) = \{2, 8\}$
$R^1{}_A(3) = \{3\}$
$R^1{}_A(4) = \{4, 8\}$
$R^1{}_A(5) = \{4, 5\}$
$R^1{}_A(6) = \{6, 8\}$
$R^1{}_A(7) = \{7\}$
$R^1{}_A(8) = \{2, 4, 6, 8\}$

Table 3.2: R^1 Characteristic sets for Table 2.1 incomplete data set

R^1 Characteristic relation for Table 2.1 data set, where B = A is defined as :

$R^1(A) = \{(1, 1), (2, 2), (2, 8), (3, 3), (4, 4), (4, 8), (5, 4), (5, 5), (6, 6), (6, 8), (7, 7), (8, 2), (8, 4), (8, 6), (8, 8)\}$.

A concept A-lower and A-upper approximations of X are defined as follows:

$\underline{A}X = \cup\{R^1{}_A(x) \mid x \in X, R^1{}_A(x) \subseteq X\}$,

$\overline{A}X = \cup\{R^1{}_A(x) \mid x \in X, R^1{}_A(x) \cap X \neq \emptyset \} = \cup\{R^1{}_A(x) \mid x \in X\}$.

For incomplete data set presented in Table 2.1, the concept A-lower and A-upper approximations of the two concepts are:

$\underline{A}\{1, 2, 4, 8\} = \{1, 2, 8, 4\}$, $\underline{A}\{3, 5, 6, 7\} = \{3, 7\}$,

$\overline{A}\{1, 2, 4, 8\} = \{1, 8, 2, 4, 6\}$, $\overline{A}\{3, 5, 6, 7\} = \{3, 4, 5, 6, 8, 7\}$.

$R^1{}_A$ is a binary relation that is reflexive, but not symmetric and transitive. It is more restrictive compared to R_A

R^2characteristic relation is defined as follows:

$R^2(B) = \{ (x, y) \in UxU \mid (\forall a \in C, a(x) = a(y) \vee a(x) = * \vee a(y) = *) \}$

R^2 characteristic sets for Table 2.1, where B = A, are presented in Table 3.3.

$R^2{}_A(x)$
$R^2{}_A(1) = \{1, 3, 8\}$
$R^2{}_A(2) = \{2, 8\}$
$R^2{}_A(3) = \{1, 3\}$
$R^2{}_A(4) = \{4, 5, 8\}$
$R^2{}_A(5) = \{4, 5, 8\}$
$R^2{}_A(6) = \{6, 8\}$
$R^2{}_A(7) = \{7\}$
$R^2{}_A(8) = \{1, 2, 4, 5, 6, 8\}$

Table 3.3: R^2 Characteristic sets for Table 2.1 incomplete data set

R^2 Characteristic relation for Table 2.1, where B = A is defined as follows :

$R^2(A) = \{(1, 1), (1, 3), (1, 8), (2, 2), (2, 8), (3, 1), (3, 3), (4, 4), (4, 5), (4, 8), (5, 4), (5, 5), (5, 8), (6, 6), (6, 8), (7, 7), (8, 1), (8, 2), (8, 4), (8, 5), (8, 6), (8, 8)\}$.

A concept A-lower and A-upper approximations of X are defined as follows:

$\underline{A}X = \cup\{R^2{}_A(x) \mid x \in X, R^2{}_A(x) \subseteq X\}$

$\overline{A}X = \cup\{R^2{}_A(x) \mid x \in X, R^2{}_A(x) \cap X \neq \emptyset \} = \cup\{R^2{}_A(x) \mid x \in X\}$

For incomplete data set presented in Table 2.1, the concept A-lower and A-upper approximations of the two concepts are:

$\underline{A}\{1, 2, 4, 8\} = \{2, 8\}$, $\underline{A}\{3, 5, 6, 7\} = \{7\}$,

$\overline{A}\{1, 2, 4, 8\} = \{1, 3, 8, 2, 4, 5, 6\}$, $\overline{A}\{3, 5, 6, 7\} = \{1, 3, 4, 5, 8, 6, 7\}$.

$R^2{}_A$ is a binary relation that is reflexive. It is less restrictive compared to R_A .

R^3 characteristic relation for Table 2.1, where B = A is defined as follows:

$R^3(B) =\{ (x, y) \in UxU \mid (\forall a \in C, a(x) = a(y) \vee a(x)= * \vee a(y)= *) \wedge (\forall a \in D, a(x) = a(y)) \}$

R^3 characteristic sets for Table 2.1, where B = A , are presented in Table 3.4.

$R^3{}_A(x)$
$R^3{}_A(1) = \{1, 3\}$
$R^3{}_A(2) = \{2, 8\}$
$R^3{}_A(3) = \{1, 3\}$
$R^3{}_A(4) = \{4, 5, 8\}$
$R^3{}_A(5) = \{4, 5\}$
$R^3{}_A(6) = \{6, 8\}$
$R^3{}_A(7) = \{7\}$
$R^3{}_A(8) = \{2, 4, 6, 8\}$

Table 3.4: R^3 characteristic sets for Table 2.1.

R^3 characteristic relation for Table 2.1, where B = A , is defined as follows :

$R^3(A) = \{(1, 1), (1, 3), (2, 2), (2, 8), (3, 1), (3, 3), (4, 4), (4, 5), (4, 8), (5, 4), (5, 5), (6, 6), (6, 8), (7, 7), (8, 2), (8, 4), (8, 6), (8, 8)\}$.

A concept A-lower and A-upper approximations of X are defined as follows:

$\underline{A}X = \cup\{R^3{}_A(x) \mid x \in X, R^3{}_A(x) \subseteq X\}$,

$\overline{A}X = \cup\{R^3{}_A(x) \mid x \in X, R^3{}_A(x) \cap X \neq \emptyset \} = \cup\{R^3{}_A(x) \mid x \in X\}$.

For incomplete data set presented in Table 2.1, the concept A-lower and A-upper approximations of the two concepts are:

$\underline{A}\{1, 2, 4, 8\} = \{2, 8\}$, $\underline{A}\{3, 5, 6, 7\} = \{7\}$,

$\overline{A}\{1, 2, 4, 8\} = \{1, 3, 2, 8, 4, 5, 6\}$, $\overline{A}\{3, 5, 6, 7\} = \{1, 3, 4, 5, 6, 8, 7\}$.

3.4 Maximal Consistent Blocks

The concept of maximal consistent block was introduced in [2]. A maximal consistent block is defined as the maximal collection of objects in which all objects are similar, and are indiscernible based on the attribute values available. Let K(A) be a characteristic relation and $K_A(x)$ be a characteristic set, we say that $Y \in K_A(x)$ is a maximal characteristic set if and only if Y is a maximal subset of $K_A(x)$, such that for any $x, y \in Y$, (x, y) or $(y, x) \in K(A)$. For $x \in U$, maximal characteristic neighborhood system is derived for each characteristic set K(A) consisting of its maximal subsets, and it is defined as :

$\mathrm{NS}(\mathrm{K}(\mathrm{A})) = \{\mathrm{Y} \subset \mathrm{K_A}(\mathrm{x}) : \mathrm{Y}$ is a maximal set of $\mathrm{K_A}(\mathrm{x})\ \}$

The neighborhood maximal consistent block sets for Table 2.1, are presented in Table 3.5.

$\mathrm{NS}(\mathrm{K_A}(\mathrm{x}))$
$\mathrm{NS}(\mathrm{K_A}(1)) = \{\{1\}\}$
$\mathrm{NS}(\mathrm{K_A}(2)) = \{\{2, 8\}\}$
$\mathrm{NS}(\mathrm{K_A}(3)) = \{\{3\}\}$
$\mathrm{NS}(\mathrm{K_A}(4)) =\{\{4, 8\}\}$
$\mathrm{NS}(\mathrm{K_A}(5)) = \{\{5\}\}$
$\mathrm{NS}(\mathrm{K_A}(6)) = \{\{6, 8\}\}$
$\mathrm{NS}(\mathrm{K_A}(7)) = \{\{7\}\}$
$\mathrm{NS}(\mathrm{K_A}(8)) = \{\{2, 8\}, \{4, 8\}, \{6, 8\}\}$

Table 3.5: Maximal consistent block Table 2.1 incomplete data set

The maximal consistent block relation for Table 2.1, are defined as follows :

$\mathrm{R}(\mathrm{A}) = \{(1,1), (2, 2), (2, 8), (3, 3), (4, 4), (4, 8), (5, 5), (6, 6), (6, 8), (7, 7), (8, 2), (8, 4), (8, 6), (8, 8)\}$.

A concept A-lower and A-upper approximations of X are defined as follows:

$\underline{A}\mathrm{X} = \cup\{\mathrm{Y} \in \mathrm{NS}((\mathrm{K_A}(\mathrm{x})):\mathrm{x} \in \mathrm{X},\mathrm{Y} \subseteq \mathrm{X}\}$,

$\overline{A}\mathrm{X} = \cup\{\mathrm{Y} \in \mathrm{NS}((\mathrm{K_A}(\mathrm{x})):\mathrm{x} \in \mathrm{X},\mathrm{Y} \cap \mathrm{X} \neq \emptyset\ \}$.

For incomplete data set presented in Table 2.1, the concept A-lower and A-upper approximations of the two concepts are:

$\underline{\mathrm{A}}\{1, 2, 4, 8\} = \{1, 8, 2, 4, 6\}$, $\underline{\mathrm{A}}\{3, 5, 6, 7\} = \{3, 4, 5, 8, 6, 7\}$,

$\overline{A}\{1, 2, 4, 8\} = \{1, 8, 2, 4, 6\}$, $\overline{A}\{3, 5, 6, 7\} = \{3, 4, 5, 8, 6, 7\}$.

3.5 Analysis of Modified Relations

Based on results of incomplete data set in Table 2.1, the analysis of modified relation is presented as follows :

- $R^1{}_A$ will ignore the similarity in objects that only have "?" and "*" interpretation. Two objects are considered similar only if all of their known attribute values are equal. For example, the characteristics set of object 1 is $\{1, 8\}$ and it considers objects 1 and 8 to be similar, but $R^1{}_A$ treats each object as a different since there is no known equal attribute value, $R^1{}_A(1) = \{4, 5\}$. As well for object 5, the characteristics set is $\{4, 5, 8\}$, but since object 5 and 8 have no know equal attribute value, $R^1{}_A(5)= \{4, 5\}$.

- $R^2{}_A$ will ignore the similarity in the objects that only have "?" interpretation. Two objects are considered similar only if there exits at least one known equal attribute value or missing attribute value is interpreted as "*". For example, the characteristics set of object 1 is $\{1, 8\}$, and it considers objects 1 and 8 to be similar, however $R^2{}_A$ treats objects 1, 3 and 8 to be similar, since objects 1 and 3 have attribute value Nausea to be equal and object 8 have missing attribute value Nausea interpretation to be " *". $R^1{}_A(1)= \{1, 3, 8\}$.

- R' and $R^3{}_A$ will ignore the similarity in the objects that only have "?" and "*" interpretation. Two objects are considered similar only if there exits at least one known equal attribute value or missing attribute value is interpreted as " * ". For example, the characteristics set of object 1 is $\{1, 8\}$ and it considers objects 1 and 8 to be similar, but R' and R3 treat each object as different since there is no known equal attribute value and Headache attribute value is not "*", however objects 1 and 3 are consider to be similar since attribute value Nausea is equal. $R'_A(1)$ and $R^3{}_A(1)= \{1, 3\}$.

- The maximal consistent block ignores the similarity in the objects, which has "?" interpretation. The characteristics set of object 5 is $\{4, 5, 8\}$ considers objects 4,5 and 8 to be similar, but maximal consistent block treats each object as a separate block. Maximal consist block of object 5 is $\{5\}$.

- Maximal consistent block provides better discernibility for the objects with "*" interpretation. For example, the characteristics set of objects 8 considers objects $\{2, 4, 6, 8\}$ to be similar, but the maximal consistent block of object 8 splits them into three separate blocks {{2, 8}, {4, 8}, {6, 8}}.

3.6 Local Definability of Modified Relations

The author of this paper [5] studied the modifications explained above that is suggested to characteristic relation in term of local definability to validate whether the modified relations are suitable for data mining.

- R' characteristic relation where characteristic class $R'_A(x)$, $x \in U$ is proven to be not ***B-locally definable*** by showing that for Table 2.1 $R'_A(1) = \{1, 3\}$ and $R'_A(3) = \{1, 3\}$, while attribute-value blocks for Table 2.1 show that whenever there is a object 1, object 8 must exist in the same block. This means that any intersection of blocks containing object 1, it must includes object 8 as well, however $8 \notin R'_A(1)$, $R'_A(3)$.
- R^1_A characteristic relation where characteristic class $R^1_A(x)$, $x \in U$ is not ***B-locally definable*** by showing that for Table 2.1 $R^1_A(1) = \{1\}$, while attribute-value blocks for Table 2.1 shows that any intersection of blocks containing object 1, it must include object 8 as well, however $8 \notin R^1_A(1)$.
- R^2_A characteristic relation where characteristic class $R^2_A(x)$, $x \in U$ is not ***B-locally definable*** by showing that for Table 2.1 $R^2_A(3) = \{1, 3\}$, while attribute-value blocks for Table 2.1 shows that any intersection of blocks containing object 1, it must include object 8 as well, however $8 \notin R^2_A(3)$.
- R^3_A characteristic relation where characteristic class $R^3_A(x)$, $x \in U$ is not ***B-locally definable*** by showing that for Table 2.1 $R^3_A(3) = \{1, 3\}$, while attribute-value blocks for Table 2.1 shows that any intersection of blocks containing object 1, it must include object 8 as well, however $8 \notin R^3$.

- Maximal consistent blocks are not ***B-locally definable*** when missing attribute values are of "do not care" and lost conditions. NS(K_A(1)) = {{1}}, while attribute-value blocks for Table 2.1 shows that any intersection of blocks containing object 1, it must include object 8 as well, but 8 $\notin$ NS(K_A(1)) = {{1}} , however maximal consistent block is ***B-locally definable*** when missing attribute values of do not care condition only. For example incomplete data set Table 3.6 maximal consistent blocks of A are [[1, 3], [1, 8], [2, 8], [4, 5, 8], [6, 8], [7]] and attribute-value blocks and characteristic sets are presented respectively in Tables 3.7 and 3.8. X is a maximal consistent block of B if and only if X is an intersection of all characteristic sets K_A(x) where x $\in$ X. For example maximal consistent block [1,8] is represented as K_B(1) $\cap$ K_B(8) = {1, 3, 8} $\cap$ {1, 2, 4, 5, 6, 8} = {1, 8}. Each characteristic set K_A(x) is ***B-globally definable*** and can be presented as an intersection of some blocks of attribute-value pairs with specified attribute values. Let us denote the set of such attribute-value pairs by T_x. For any maximal consistent block X there exists x $\in$ U such that X $\subseteq$ K_A(x). For any y $\in$ X, if a(x) is specified then either A(y) = a(x) or a(y) = "*", [1,8] is represented as [(Temperature, High)] $\cap$ [(Nausea, no)] $\cap$ [(Headache, yes)] = {1, 3, 4, 5, 8} $\cap$ {1, 3, 6, 8} $\cap$ {1, 2, 4, 5, 6, 8} = {1, 8}. Maximal consistent block X can be presented as an intersection of blocks of attribute-value pairs from T_x. Therefore X is B-locally definable. In general, maximal consistent block is not ***B-globally definable***. For example set {1,8} is not ***B-globally definable***, but it is ***B-locally definable***.

case	Temperature	Headache	Nausea	Flu
1	High	*	No	Yes
2	Very-high	Yes	Yes	Yes
3	*	No	No	No
4	High	Yes	Yes	Yes
5	High	*	Yes	No
6	Normal	yes	No	No
7	Normal	No	Yes	No
8	*	Yes	*	Yes

Table 3.6: An example of incomplete decision table

(a,v)	[(a,v)]
(Temperature, High)	$\{1, 3, 4, 5, 8\}$
(Temperature, Very-high)	$\{2, 3, 8\}$
(Temperature, normal)	$\{3, 6, 7, 8\}$
(Headache, yes)	$\{1, 2, 4, 5, 6, 8\}$
(Headache, no)	$\{1, 3, 5, 7\}$
(Nausea, no)	$\{1, 3, 6, 8\}$
(Nausea, yes)	$\{2, 4, 5, 7, 8\}$

Table 3.7: Attribute-value blocks for Table 3.6.

$K_B(x)$
$K_B(1) = \{1, 3, 8\}$
$K_B(2) = \{2, 8\}$
$K_B(3) = \{1, 3\}$
$K_B(4) = \{4, 5, 8\}$
$K_B(5) = \{4, 5, 8\}$
$K_B(6) = \{6, 8\}$
$K_B(7) = \{7\}$
$K_B(8) = \{1, 2, 4, 5, 6, 8\}$

Table 3.8: Characteristic sets for Table 3.6.

Chapter 4

Experiments

The objective of this thesis is to investigate the impact of the modifications suggested to characteristic relation on concept lower and upper approximations and rules induced in a comparison to the characteristic relation. An experimental comparative analysis is conducted between characteristic sets and modifications sets in terms of cardinalities of lower and upper approximations of each concept and decision rules induced by each modification. We conducted experiments on four incomplete data sets where each data set has different interpretations and percentage of missing data. Each data set have three interpretations of missing data, where missing attribute values will only be lost conditions, "do not care" conditions or both types of missing attribute values. Brief description of the input data sets used in experiments is shown in Section 4.1. Results of comparison between upper and lower Approximations of each modification are shown in Section 4.2. Section 4.3 shows experiments on rules induced from different relation sets.

4.1 Data Sets

Experiments were conducted on the following incomplete data sets with varying levels of missing data are shown in Table 4.1. Below is a brief description for all the data sets.

- Breast Cancer Data set :

  This data set is obtained from the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia. There are 9 attributes and total of 277 instances. The attributes are categorical. Dataset has 2 decision classes, which specify the condition as "recurrence" or "non-recurrence" events.

- Hepatitis Data set:

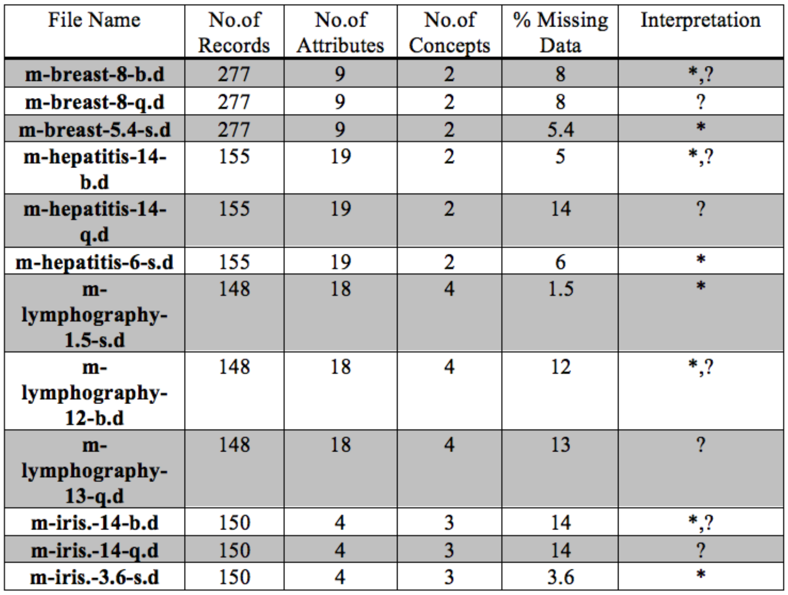
  It has 19 attributes and total of 155 instances. There are 2 decision classes with each labeled as one of the two types possible either "no" or "yes". All attributes are categorical.

- Lymph Data set:

  The lymphography domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. There are 18 attributes and 148 instances in the dataset.All attributes are categorical. There are 4 decision classes, with each labeled as "one", "two", "three" or "four".

- Iris Data set:

  This is the best-known database to be found in the pattern recognition literature. There are four attributes and 150 records. There are 3 classes labelled as "Iris Setosa", "Iris Versicolor" and "Iris Virginica".



File Name	No.of Records	No.of Attributes	No.of Concepts	% Missing Data	Interpretation
m-breast-8-b.d	277	9	2	8	*,?
m-breast-8-q.d	277	9	2	8	?
m-breast-5.4-s.d	277	9	2	5.4	*
m-hepatitis-14-b.d	155	19	2	5	*,?
m-hepatitis-14-q.d	155	19	2	14	?
m-hepatitis-6-s.d	155	19	2	6	*
m-lymphography-1.5-s.d	148	18	4	1.5	*
m-lymphography-12-b.d	148	18	4	12	*,?
m-lymphography-13-q.d	148	18	4	13	?
m-iris.-14-b.d	150	4	3	14	*,?
m-iris.-14-q.d	150	4	3	14	?
m-iris.-3.6-s.d	150	4	3	3.6	*

Table 4.1: Incomplete data sets used in experiments

4.2 Results of Comparison Between Upper and Lower Approximations

Experiments were conducted on data sets listed in Section 4.1 with varying levels of missing attribute values. The cardinalities of concept lower and upper approximations of characteristics sets and each new modified sets are compared. Results of a comparison between upper and lower approximations are shown in Tables 4.2, 4.3, 4.4 and 4.5. Data sets approximations cardinalities are exhibiting the following results:

- The cardinality of upper approximation using R' characteristic relation sets for incomplete data sets is greater than or equal to the cardinality of upper approximation of characteristic sets, while the lower approximation cardinality is smaller than lower approximation cardinality of characteristic sets.

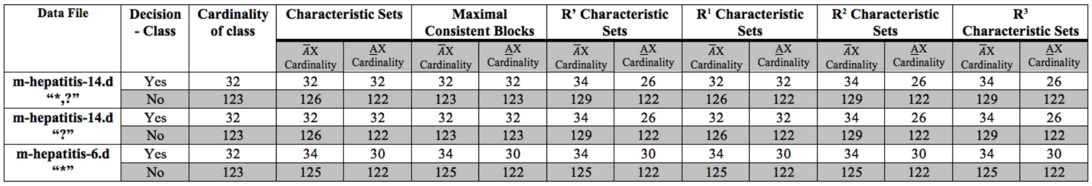
- The cardinality of upper approximation using R^1 characteristic relation sets for incomplete data sets is either smaller than or equal to the cardinality of upper approximation of characteristic sets. The cardinality of lower approximation using R^1 is greater than or equal to the cardinality of lower approximation using characteristic sets.

- The cardinality of upper approximation using R^2 for incomplete data sets is greater than or equal to the cardinality of upper approximation using characteristic sets, while the lower approximation cardinality is smaller than or equal to the cardinality of lower approximation using characteristic sets.

- The cardinality of upper approximation using R^3 characteristic relation sets for incomplete data sets is either greater than or equal to the cardinality of upper approximation using characteristic sets, while the lower approximations is smaller than or equal to the lower approximation cardinality of characteristic set.

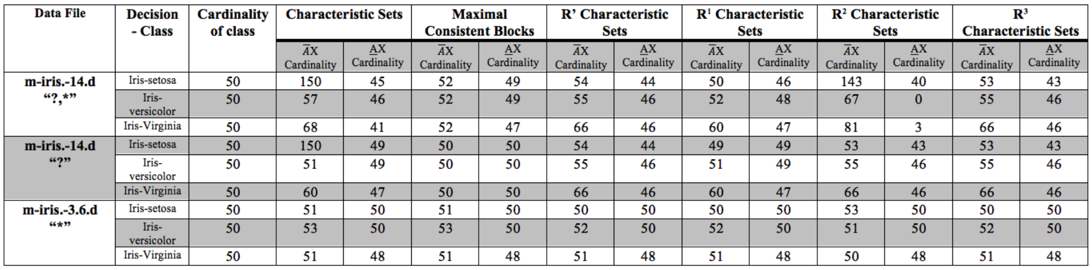
- The cardinality of upper approximation using maximal consistent blocks for incomplete data sets is smaller than or equal to the cardinality of upper approximation of characteristic sets. The cardinality of lower approximation using maximal consistent blocks for incomplete data sets is either greater than or equal to the cardinality of lower approximation using characteristic sets.

- The variation between approximations cardinalities for characteristics sets and modification sets is larger for incomplete data set with "?" interpretation, when compared to incomplete data set with both missing attribute value interpretations, and it is observed to be the smallest for incomplete data sets with "*" interpretation.



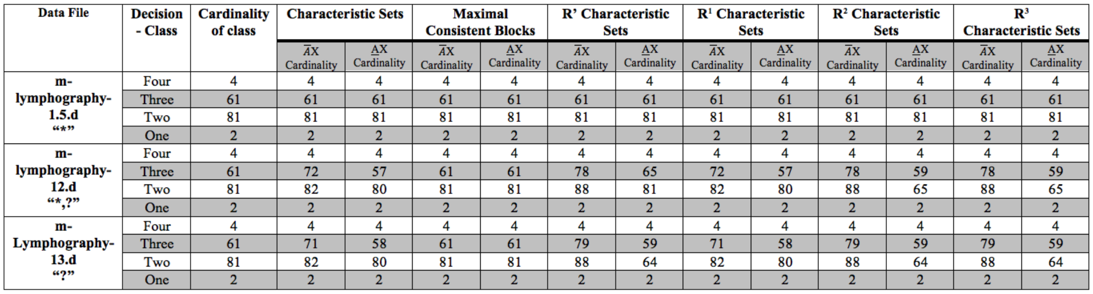
Data File	Decision - Class	Cardinality of class	Characteristic Sets		Maximal Consistent Blocks		R’ Characteristic Sets		R^1 Characteristic Sets		R^2 Characteristic Sets		R^3 Characteristic Sets	
			$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality
m-hepatitis-14.d “*,?”	Yes	32	32	32	32	32	34	26	32	32	34	26	34	26
	No	123	126	122	123	123	129	122	126	122	129	122	129	122
m-hepatitis-14.d “?”	Yes	32	32	32	32	32	34	26	32	32	34	26	34	26
	No	123	126	122	123	123	129	122	126	122	129	122	129	122
m-hepatitis-6.d “*”	Yes	32	34	30	34	30	34	30	34	30	34	30	34	30
	No	123	125	122	125	122	125	122	125	122	125	122	125	122

Table 4.2: Approximation cardinality for data set hepatitis.d



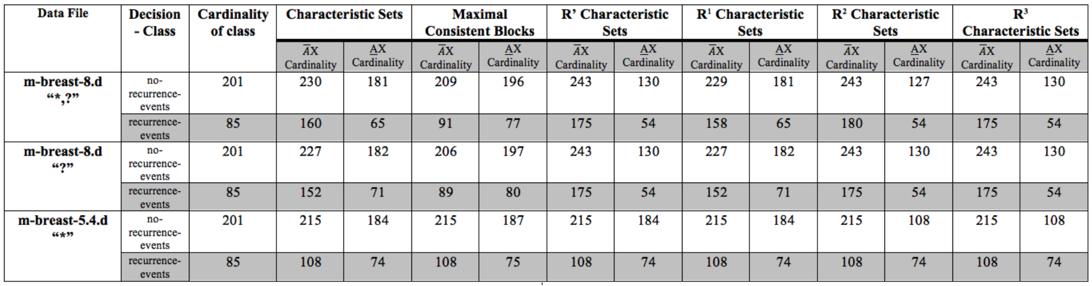
Data File	Decision - Class	Cardinality of class	Characteristic Sets		Maximal Consistent Blocks		R’ Characteristic Sets		R^1 Characteristic Sets		R^2 Characteristic Sets		R^3 Characteristic Sets	
			$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality
m-iris.-14.d “?,*”	Iris-setosa	50	150	45	52	49	54	44	50	46	143	40	53	43
	Iris-versicolor	50	57	46	52	49	55	46	52	48	67	0	55	46
	Iris-Virginia	50	68	41	52	47	66	46	60	47	81	3	66	46
m-iris.-14.d “?”	Iris-setosa	50	150	49	50	50	54	44	49	49	53	43	53	43
	Iris-versicolor	50	51	49	50	50	55	46	51	49	55	46	55	46
	Iris-Virginia	50	60	47	50	50	66	46	60	47	66	46	66	46
m-iris.-3.6.d “*”	Iris-setosa	50	51	50	51	50	50	50	50	50	53	50	50	50
	Iris-versicolor	50	53	50	53	50	52	50	52	50	51	50	52	50
	Iris-Virginia	50	51	48	51	48	51	48	51	48	50	48	51	48

Table 4.3: Approximation cardinality for data set iris.d



Data File	Decision - Class	Cardinality of class	Characteristic Sets		Maximal Consistent Blocks		R’ Characteristic Sets		R^1 Characteristic Sets		R^2 Characteristic Sets		R^3 Characteristic Sets	
			$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality
m-lymphography-1.5.d “*”	Four	4	4	4	4	4	4	4	4	4	4	4	4	4
	Three	61	61	61	61	61	61	61	61	61	61	61	61	61
	Two	81	81	81	81	81	81	81	81	81	81	81	81	81
	One	2	2	2	2	2	2	2	2	2	2	2	2	2
m-lymphography-12.d “*,?”	Four	4	4	4	4	4	4	4	4	4	4	4	4	4
	Three	61	72	57	61	61	78	65	72	57	78	59	78	59
	Two	81	82	80	81	81	88	81	82	80	88	65	88	65
	One	2	2	2	2	2	2	2	2	2	2	2	2	2
m-Lymphography-13.d “?”	Four	4	4	4	4	4	4	4	4	4	4	4	4	4
	Three	61	71	58	61	61	79	59	71	58	79	59	79	59
	Two	81	82	80	81	81	88	64	82	80	88	64	88	64
	One	2	2	2	2	2	2	2	2	2	2	2	2	2

Table 4.4: Approximation cardinality for data set lymphography.d



Data File	Decision - Class	Cardinality of class	Characteristic Sets		Maximal Consistent Blocks		R’ Characteristic Sets		R^1 Characteristic Sets		R^2 Characteristic Sets		R^3 Characteristic Sets	
			$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality	$\overline{A}X$ Cardinality	$\underline{A}X$ Cardinality
m-breast-8.d “*,?”	no-recurrence-events	201	230	181	209	196	243	130	229	181	243	127	243	130
	recurrence-events	85	160	65	91	77	175	54	158	65	180	54	175	54
m-breast-8.d “?”	no-recurrence-events	201	227	182	206	197	243	130	227	182	243	130	243	130
	recurrence-events	85	152	71	89	80	175	54	152	71	175	54	175	54
m-breast-5.4.d “*”	no-recurrence-events	201	215	184	215	187	215	184	215	184	215	108	215	108
	recurrence-events	85	108	74	108	75	108	74	108	74	108	74	108	74

Table 4.5: Approximation cardinality for data set breast.d

4.3 Experiments on Rules Induced from Modified Characteristic Relations

This section will investigate the impact of the modifications suggested to characteristic relation on rule induced. Experiments are conducted on rule induced from different characteristic relations implantations of incomplete data sets. The data set from which rules are induced is called training set, and the data set that the accuracy of classification with the induced rules is measured is called testing set. The LERS classification method used in classifying unseen cases, cross validation and error rate are explained in subsection 4.2.1. Subsection 4.2.2 will describe the experiment process of calculating classification error rate and results of the experiments are presented in Subsection 4.2.3.

4.3.1 LERS Classification System and Cross Validation

Classification method used of LER learning system introduced in [14], which is the classification of an unseen case to concept depend on three parameters: strength, support and partial or complete matching factor. Strength is the total number of cases of training data sets that are correctly classified by the rule, and specificity is the total number of conditions in rule r. The support for concept C is defined as:

$$\mathrm{C}_{\mathrm{Support}} = \sum_{Complety\ matching\ rule\ r\ with\ concept\ C} Strength\ (r) * specifity\ (r)$$

If complete matching for case x is not achieved, search for partially matched rules with case x. Partially matched factor (pmf) is that attributes values of a case x that match at least one of the conditions of rule r, and defined as the number of conditions matched by the total number of conditions of rule r. When there is a partially matched cases the support of concept C is defined as :

$$C_{\text{Support}} = \sum_{partially\ matching\ rule\ r\ with\ concept\ C} pmf(r) * \ Strength\ (r) * specifity\ (r)$$

If rules r_1, r_2,...., r_n all completely or partially match a testing data case x and all rules have the same concept C, case x will be classified as concept C. If the matching rules refer to different concepts, then support is calculated for every concept. The concept with the largest support will be concept that the case x is classified to.

Ten-fold cross validation is used to measure the accuracy of classification model produced from the rule induced of the incomplete data set. The incomplete data set is randomly reordered, then the reordered data set is partitioned into ten subsets, where each subset contain roughly 10% of the data set. Each subset in the ten subsets is used as testing data, while the other subsets are used as training data. The rule set induced from the training set is used to classify every case of the testing set. The error rate is measured as total number of incorrectly classified or unclassified cases by the total number of cases. Ten runs of rule induction and classification are conducted and the average error rate is used to measure the accuracy of classification of each modification.

4.3.2 Experimental Procedure

LEM2 is implemented for rule induction where the input to LEM2 is the upper and lower concept approximations of each relation sets. The procedure of the experiment is shown in Figure 4.1. For each incomplete data set, LEM2 was applied and the algorithm will produce two sets of rules : possible and certain rule sets. Average error rate of the classification was calculated for each rule sets.

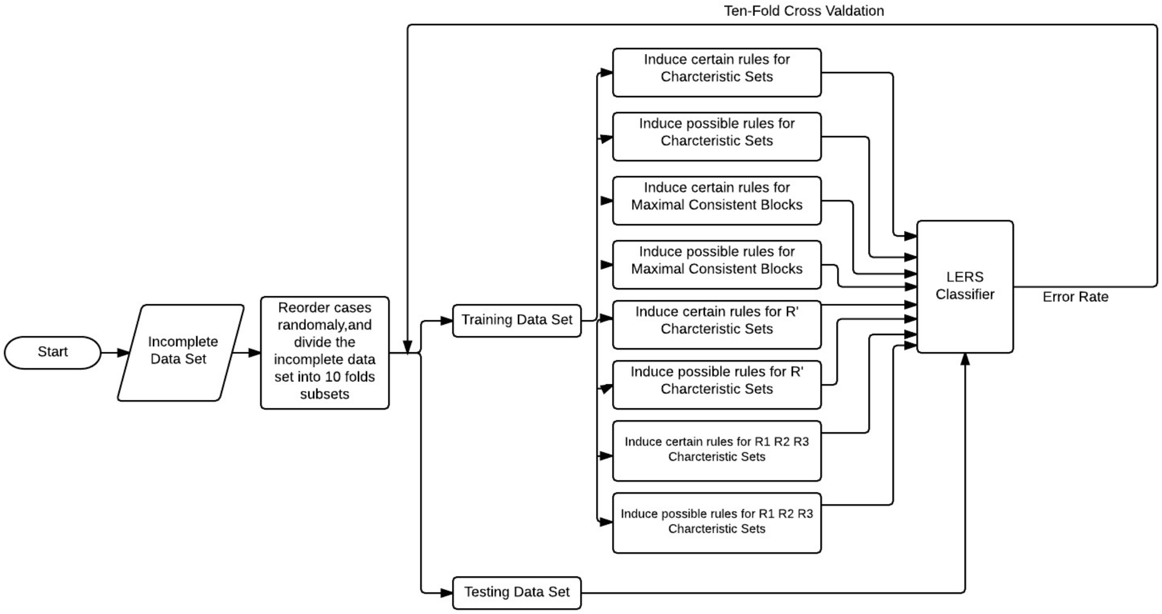
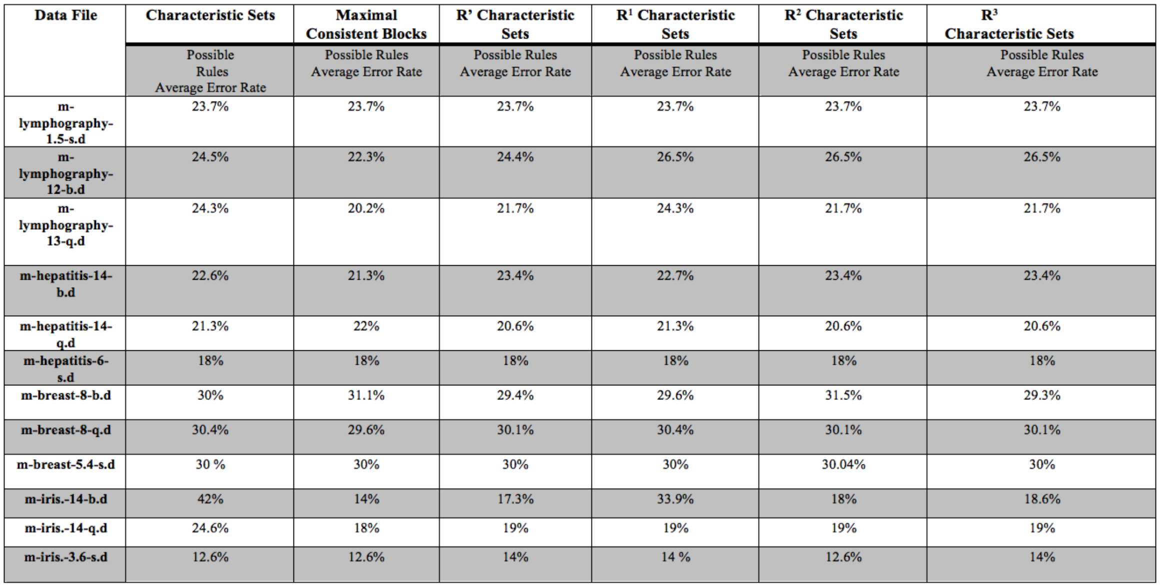

Ten-Fold Cross Valdation
Start
Incomplete Data Set
Reorder cases randomaly,and divide the incomplete data set into 10 folds subsets
Training Data Set
Testing Data Set
Induce certain rules for Charcteristic Sets
Induce possible rules for Charcteristic Sets
Induce certain rules for Maximal Consistent Blocks
Induce possible rules for Maximal Consistent Blocks
Induce certain rules for R' Charcteristic Sets
Induce possible rules for R' Charcteristic Sets
Induce certain rules for R1 R2 R3 Charcteristic Sets
Induce possible rules for R1 R2 R3 Charcteristic Sets
LERS Classifier
Error Rate

Figure 4.1: Experimental procedure

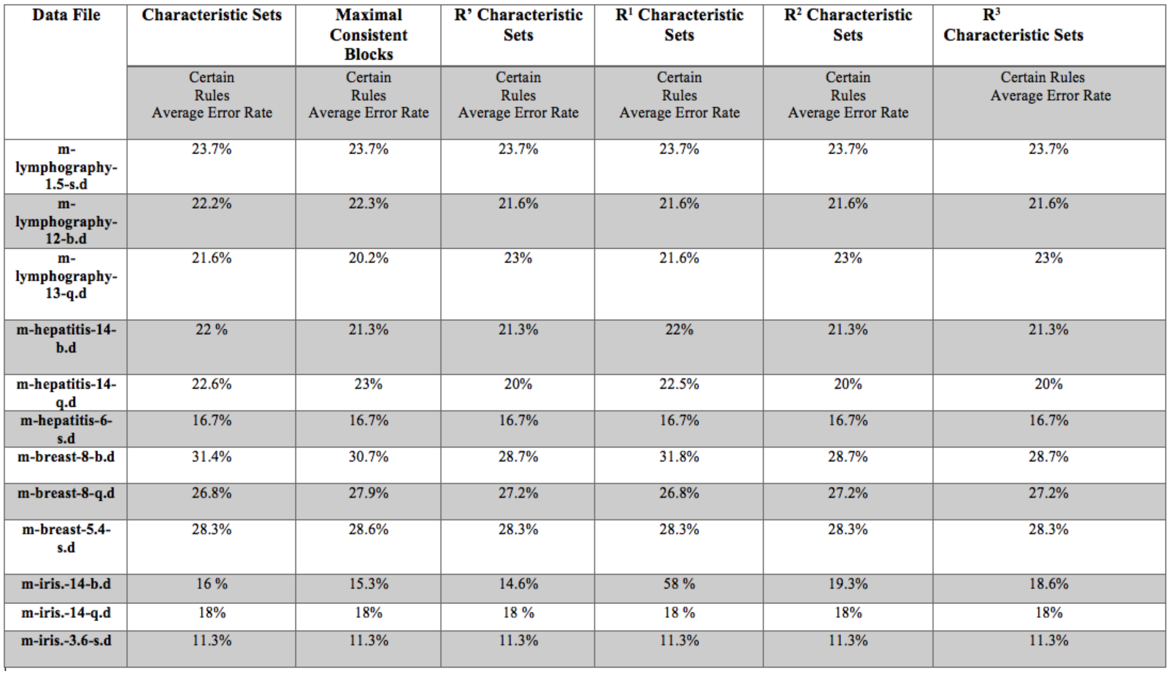
4.3.3 Experimental Results

The average error rates of ten-fold cross validation for incomplete data sets are shown in Table 4.6 and Table 4.7. Our result shows that there is no significant difference between the qualities of rule induced from each modification except for the incomplete data set iris with interpretation "?", where the possible average error rate is much higher compared to all other modifications. It is noted that possible rules average error rate of R^1 and maximal consistent block for most incomplete data sets is slightly smaller than or equal to characteristic set possible rules average error rate, while certain rules average error rate of R', R^2 and R^3 for most incomplete data sets is slightly smaller than or equal to characteristic set certain rules average error rate. Certain rules have smaller average error rate compared to possible rules average error rate.



Data File	Characteristic Sets	Maximal Consistent Blocks	R’ Characteristic Sets	R^1 Characteristic Sets	R^2 Characteristic Sets	R^3 Characteristic Sets
	Possible Rules Average Error Rate	Possible Rules Average Error Rate	Possible Rules Average Error Rate	Possible Rules Average Error Rate	Possible Rules Average Error Rate	Possible Rules Average Error Rate
m-lymphography-1.5-s.d	23.7%	23.7%	23.7%	23.7%	23.7%	23.7%
m-lymphography-12-b.d	24.5%	22.3%	24.4%	26.5%	26.5%	26.5%
m-lymphography-13-q.d	24.3%	20.2%	21.7%	24.3%	21.7%	21.7%
m-hepatitis-14-b.d	22.6%	21.3%	23.4%	22.7%	23.4%	23.4%
m-hepatitis-14-q.d	21.3%	22%	20.6%	21.3%	20.6%	20.6%
m-hepatitis-6-s.d	18%	18%	18%	18%	18%	18%
m-breast-8-b.d	30%	31.1%	29.4%	29.6%	31.5%	29.3%
m-breast-8-q.d	30.4%	29.6%	30.1%	30.4%	30.1%	30.1%
m-breast-5.4-s.d	30 %	30%	30%	30%	30.04%	30%
m-iris.-14-b.d	42%	14%	17.3%	33.9%	18%	18.6%
m-iris.-14-q.d	24.6%	18%	19%	19%	19%	19%
m-iris.-3.6-s.d	12.6%	12.6%	14%	14 %	12.6%	14%

Table 4.6: Possible rule average error rate



Data File	Characteristic Sets	Maximal Consistent Blocks	R’ Characteristic Sets	R^1 Characteristic Sets	R^2 Characteristic Sets	R^3 Characteristic Sets
	Certain Rules Average Error Rate	Certain Rules Average Error Rate	Certain Rules Average Error Rate	Certain Rules Average Error Rate	Certain Rules Average Error Rate	Certain Rules Average Error Rate
m-lymphography-1.5-s.d	23.7%	23.7%	23.7%	23.7%	23.7%	23.7%
m-lymphography-12-b.d	22.2%	22.3%	21.6%	21.6%	21.6%	21.6%
m-lymphography-13-q.d	21.6%	20.2%	23%	21.6%	23%	23%
m-hepatitis-14-b.d	22 %	21.3%	21.3%	22%	21.3%	21.3%
m-hepatitis-14-q.d	22.6%	23%	20%	22.5%	20%	20%
m-hepatitis-6-s.d	16.7%	16.7%	16.7%	16.7%	16.7%	16.7%
m-breast-8-b.d	31.4%	30.7%	28.7%	31.8%	28.7%	28.7%
m-breast-8-q.d	26.8%	27.9%	27.2%	26.8%	27.2%	27.2%
m-breast-5.4-s.d	28.3%	28.6%	28.3%	28.3%	28.3%	28.3%
m-iris.-14-b.d	16 %	15.3%	14.6%	58 %	19.3%	18.6%
m-iris.-14-q.d	18%	18%	18 %	18 %	18%	18%
m-iris.-3.6-s.d	11.3%	11.3%	11.3%	11.3%	11.3%	11.3%

Table 4.7: Certain rule average error rate

Chapter 5

Conclusion

In this thesis, we investigated the validity of modifications suggested to the definition of characteristic relation defined for the incomplete data set. We showed that all modified characteristic relations are not locally definable except for maximal consistent blocks that are restricted to data set with " do not care" conditions. LEM2 algorithm was implemented to induce certain and possible rules from the incomplete data set. In our research twelve incomplete data sets with different interpretations of missing attribute values were used to conduct an experimental comparative analysis of the cardinalities of lower and upper approximations of each concept. In term of upper approximations, R^2, R^3 and R' achieve no better accuracy than characteristics sets, but R^1 and maximal consistent blocks have higher accuracy compared to characteristics sets. In term of Lower approximation, R^1 and maximal consistent blocks achieve no better accuracy than characteristics sets, but R^2, R^3 and R' achieve better accuracy compared to characteristic set. To measure the classification average error rate for induced rules, ten-fold cross validation was implemented. Our results show that possible rules induced using R^1 and maximal consistent blocks are slightly more consistent compared to characteristic set possible rules, and certain rules induced using R^2, R^3 and R' are slightly more consistent compared to characteristic set certain rules. To conclude, even though modified definitions of characteristic sets improved the accuracy and consistency of possible and certain rules induced compared to characteristic sets rules, the modified sets are not locally definable. Suggested new modified relations should not be used in data mining except for maximal consistent blocks, when the interpretation of missing attribute values is of "do not care" conditions.

References

[1] Jerzy W. Grzymala-Busse, A Rough Set Approach to Data with Missing Attribute Values, 2006, LNAI 4062, pp. 58–67.

[2] Yee Leung and Deyu Li, Maximal consistent block technique for rule acquisition in incomplete information systems, Information Sciences, 153 (2003) 85-106.

[3] Marzena Kryszkiewicz, Rough set approach to incomplete information systems, Information Sciences, 112 (1998), 39-49.

[4] Marzena Kryszkiewicz, Rules in incomplete information systems, Information Sciences, 113 (1999), 271-292.

[5] Jerzy W. Grzymala-Busse and Teresa Mroczek, Definability in Mining Incomplete Data, Procedia Computer Science, 96 (2016), 179 – 186.

[6] Zdzisław Pawlak, Rough Sets, International Journal of Computer and Information Sciences, 11 (1982), 341-356.

[7] Jerzy W. Grzymala-Busse, A Local Version of the MLEM2 Algorithm for Rule Induction, Fundamenta Informaticae, 100 (2010) 1–18.

[8] Jerzy W. Grzymala-Busse , A new version of the rule induction system LERS, Fundamenta Informaticae, 31 (1997) 27-39.

[9] Jerzy W. Grzymala-Busse, Data with missing attribute values: Generalization of indiscernibility relation and rule induction. Transactions on Rough Sets 2004,1:78-95.

[10] Grzymala-Busse JW, Rzasa W. Local and global approximations for incomplete data. Transactions on Rough Sets 2008,8:21-34.

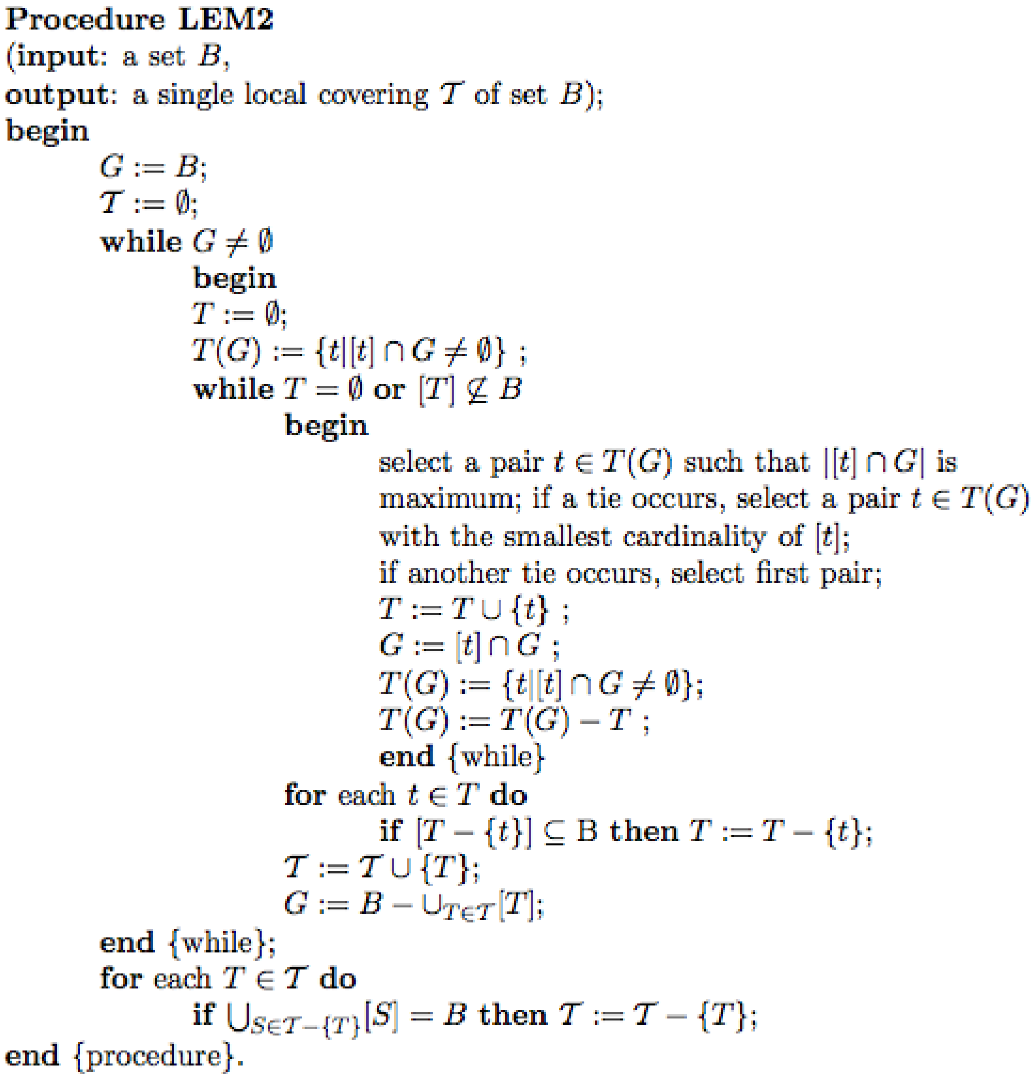
[11] Yang XB, Yang JY, Wu C, Yu DJ. Further investigation of characteristic relation in incomplete information systems. Systems Engineering -Theory & Practice 2007,27(6):155-160.

[12] Qi YS, Wei L, Sun HJ, Song YQ, Sun QS. Characteristic relations in generalized incomplete information systems. In: International Workshop on Knowledge Discovery and Data Mining, 2008. p. 519-523.

[13] Jerzy W. Grzymala-Busse,Lers-a system for learning from examples based on rough sets. In S. Roman (Ed.), Intelligent Decision Support, volume 11 of Theory and Decision Library (pp.3?18). Springer Netherlands(1992).

[14] Jerzy W. Grzymala-Busse & Wang, C. (1996). Classification and rule induction based on rough sets. In Fuzzy Systems, 1996., Proceedings of the Fifth IEEE International Conference on, volume 2 (pp. 744?747 vol.2).

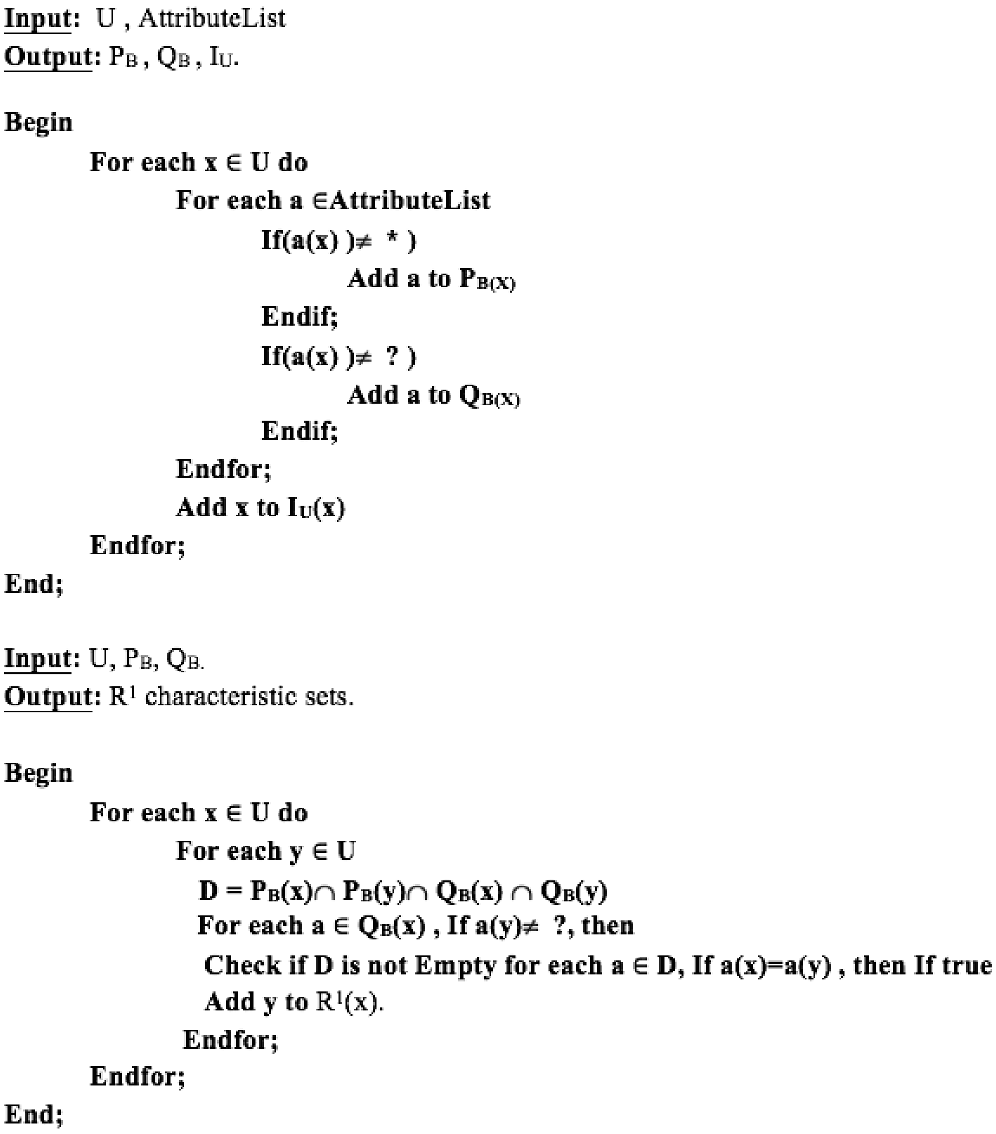
Appendix A



```
Procedure LEM2
(input: a set $B$,
output: a single local covering $\mathcal{T}$ of set $B$);
begin
    $G := B$;
    $\mathcal{T} := \emptyset$;
    while $G \neq \emptyset$
        begin
        $T := \emptyset$;
        $T(G) := \{t | [t] \cap G \neq \emptyset\}$ ;
        while $T = \emptyset$ or $[T] \not\subseteq B$
            begin
                select a pair $t \in T(G)$ such that $|[t] \cap G|$ is
                maximum; if a tie occurs, select a pair $t \in T(G)$
                with the smallest cardinality of $[t]$;
                if another tie occurs, select first pair;
                $T := T \cup \{t\}$ ;
                $G := [t] \cap G$ ;
                $T(G) := \{t | [t] \cap G \neq \emptyset\}$;
                $T(G) := T(G) - T$ ;
                end {while}
            for each $t \in T$ do
                if $[T - \{t\}] \subseteq$ B then $T := T - \{t\}$;
            $\mathcal{T} := \mathcal{T} \cup \{T\}$;
            $G := B - \cup_{T \in \mathcal{T}} [T]$;
    end {while};
    for each $T \in \mathcal{T}$ do
        if $\bigcup_{S \in \mathcal{T} - \{T\}} [S] = B$ then $\mathcal{T} := \mathcal{T} - \{T\}$;
end {procedure}.
```

Appendix B

Algorithm of computing the modifications suggested to characteristic relation is presented below.



```
Input:  U , AttributeList
Output: $P_B$ , $Q_B$ , $I_U$.

Begin
    For each $x \in U$ do
        For each $a \in$AttributeList
            If(a(x) )$\neq$ * )
                Add a to $P_{B(X)}$
            Endif;
            If(a(x) )$\neq$ ? )
                Add a to $Q_{B(X)}$
            Endif;
        Endfor;
        Add x to $I_U(x)$
    Endfor;
End;

Input: U, $P_B$, $Q_B$.
Output: $R^1$ characteristic sets.

Begin
    For each $x \in U$ do
        For each $y \in U$
          $D = P_B(x) \cap P_B(y) \cap Q_B(x) \cap Q_B(y)$
          For each $a \in Q_B(x)$ , If $a(y) \neq$ ?, then
           Check if D is not Empty for each $a \in D$, If $a(x)=a(y)$ , then If true
           Add y to $R^1(x)$.
         Endfor;
    Endfor;
End;
```

Input: U, P_B, Q_B.
Output: R^2 characteristic sets.

```
Begin
    For each $x \in U$ do
        For each $y \in U$
          Calculate $C= Q_B(x) \cap Q_B(y)$
          Check if C is not Empty, For $a \in C$
          If $a(x)=a(y)$ or $a(x)=*$ or $a(y)=*$ , If true , then
                Add y to $R^2(x)$.
        END If.
        Endfor;
    Endfor;
End;
```

Input: U, P_B, Q_B.
Output: R^3 characteristic sets.

```
Begin
    For each $x \in U$ do
        For each $y \in U$
          Calculate $C= Q_B(x) \cap Q_B(y)$
          Calculate  $D = P_B(x)\cap P_B(y)\cap Q_B(x) \cap Q_B(y)$
         Check if C and D not empty, For $a \in C$
          If $a(x)=a(y)$ or $a(x)=*$ or $a(y)=*$ , If true , then
           for each $a \in D$, If $a(x)=a(y)$ , then If true
            Add y to $R^1(x)$.
         Endfor;
    Endfor;
End;
```

Input: U, P_B, Q_B,I_U
Output: R' characteristic sets.

```
Begin
    For each $x \in U$ do
        For each $y \in U$
         Calculate $C= Q_B(x) \cap Q_B(y)$
         Calculate  $D = P_B(x) \cap P_B(y) \cap Q_B(x) \cap Q_B(y)$
         Check if C and D not empty, For $a \in C$
         If a(x)=a(y) or a(x)=* or a(y)=* , If true , then
          for each $a \in D$, If a(x)=a(y) , then If true
           Add y to $R^1(x) \cup I_U$
         Endfor;
    Endfor;
End;
```

Input: Characteristics set , Characteristics relation **R**(B)
Output: Maximal Neighborhoods Consistent Block (MNCB) for all data set objects.

```
Begin
        For each x $\in$ U do
                For each y $\in$ $K_B$(x ) do
                        If($K_B$(y).contain(x)$\neq$ true )
                                CharacteristicsSet = $K_B$(x)– y;
                        Endif;
                Endfor;
                Add to MNCB ComputeMB(CharacteristicsSet);
        Endfor;
End;
```

Input: Modified characteristics set.
Output: Maximal consistent block characteristics set of object x.

```
Begin
        If $M_B$ .contain(CharacteristicsSet) = false
                If | CharacteristicsSet | =1 then
                        Add CharacteristicsSet to $M_A$(x) and $M_B$ ;
                Else
                        If  (x, y) $\in$ R(A) for $\forall$ x, y $\in$ CharacteristicsSet then
                                Add CharacteristicsSet  to $M_A$(x) and $M_B$;
                                Return;
                        Else
                                Compute the subsets of CharacteristicsSet each of size
                                (|CharacteristicsSet | - 1), which contains x in it.
                                For each subset
                                        ComputeMB (subset);
                                Endfor;
                        Endif;
                Endif;
        Else
                Add CharacteristicsSet to $M_A$(x);
                Return;
        Endif;
End
```