On Applications of Rough Sets theory to Knowledge Discovery

Frida R. Coaquira Nina*

1. INTRODUCTION
Rough Sets theory was introduced by Z. Pawlak (1982) as a mathematical tool for data analysis. Rough sets have many applications in the field of Knowledge Discovery in Databases (KDD), such as feature selection, data reduction, discretization, etc.

When a dataset contains irrelevant (dispensable) features these can be eliminated, reducing in this way the dimension of the problem, Rough sets can be used to find subsets of relevant (indispensable) features [8]. Combining rough sets theory with a known classifier yields a wrapper method since it uses the class label information to create the indiscernability relation.

Discretization is another task in data mining applied to datasets containing continuous features. This process is done prior to the application of several KDD methods. Discretization process is an interactive method, and it is implemented based on the partitioning of the values of the attributes.

Other application of rough sets is instance selection. Equivalence relations can be found among several records of data and then some observations can be selected to form a new subset to be used in future analysis. Thus, instance selection involves extracting elementary blocks from the dataset based on an equivalence relation.

Other problem that rises in KDD is the handling of missing values. Most of the time, these values are imputed through several methods using the available data in the feature on other instances.

Rough sets theory provides tools to perform the knowledge discovery task handling the instances with missing values like the instances that behave similar to them as indicated by [7].

In this thesis we will conduct research on the application of rough sets theory in several Knowledge Discovery tasks: a search for efficient algorithms for feature selection in supervised classification based on Rough Sets theory, the development of efficient discretization methods, and imputation methods for incomplete data using Rough Sets. All the methods developed in this thesis work will be applied to datasets coming from Bioinformatics, in particular gene expression data. Gene expression data contain only continuous features, therefore it requires discretization process prior to the application of Rough sets analysis.

The vagueness of information can be seen as a property of sets imprecisely specified, whereas uncertainty can be attributed to set elements through the usage of the rough membership function, similar to the fuzzy membership function. Fuzzy methods and rough set methods are macroscopic, descriptive and numerical methods, the fuzzy methods are deductive, rough sets methods are inductive [6].

This thesis research tries to accomplish the following objectives:
1. Find efficient algorithms for feature selection in supervised classification based on rough sets.
2. Development of efficient discretization methods using Rough Sets theory.
3. Construction of imputation methods based on Rough Sets for handling incomplete data.

*Ph. D. Student, Doctoral Program in Computing and Information Sciences and Engineering, University of Puerto Rico
Mayaguez Campus, frida_cn@math.uprm.edu.
4. Find new algorithms to choose the best instances of a datasets in order to perform an efficient Knowledge Discovery on them.
5. Introduce clustering algorithms based on Rough Sets theory.
6. Evaluate the application of Rough Sets theory on Knowledge Discovery performed on real world data.

2. LITERATURE REVIEW

2.1. CLASSIFIERS

A classifier is a decision rule constructed using the available data, called the training sample, and its goal is to assign classes to instances of a new dataset, named the test sample. The features, variables of the dataset, are used to design the classifier. Examples of classifiers are: linear discriminant analysis, k-nearest neighbors classifiers, kernel classifiers, decision trees, logistic classifiers, multilayer perceptron, radial basis network, support vector machine, etc.

The proportion of future instances assigned to incorrect classes for the classifier is called the misclassification error. This error can be estimated using the training and test data sets. A combination of multiple classifiers, named an ensemble, is sometimes used to improve the performance of a single classifier. Various resampling techniques like the bootstrapping may be used to artificially create such differences. bagging and boosting are examples of ensembles.

A brief description of the classifiers to be used in this thesis follows below:

**The Linear discriminant analysis (LDA) classifier.** It is a commonly used classification method of parametric nature that uses the assumption of multivariate and equality of variance matrices. This method maximizes the ratio of between-class variance to the within-class variance in any particular dataset, thereby guaranteeing maximal separability. LDA tries to draw a decision region between the classes of the dataset using the rule: assign \( x \) to the class \( j \) that has the closest mean.

**The k-nearest neighbor (k-nn) classifier.** It finds for each instance of the dataset, the k-nearest instances using a distance measure, and the classification is decided by majority vote, with ties broken at random. If there are ties for the kth nearest neighbor, all candidates are included in the vote.

**The recursive partitioning classifier (Rpart).** It uses a tree structure to classify the instances. A classification tree will determine a set of logical if-then conditions, instead of linear equations for classifying instances. Tree classifiers use no implicit assumption about the underlying relationships between the predictor variables and the classes.

2.2 FEATURE SELECTION METHODS

The problem of feature selection consists in the search of \( d \) features from a given set of \( m \) \( (d<m) \) features, that leads to the smallest classification error. In others words, feature selection methods determine an appropriate feature subset such that the classification error is optimal [20]. The chosen features permit that pattern vectors belonging to different categories occupy compact and disjoint regions in an \( m \)-dimensional feature space.

There are two main reasons to keep the dimensionality of the features as small as possible: cost minimization and classification accuracy. Cost minimization is achieved because after the feature selection the classifiers will be faster and will use less memory [10]. A careful choice of the features is needed since a bad reduction may lead to a loss in the discrimination power and thereby a decrease in the accuracy of the resulting classifier.

The feature selection methods depend on the way that the subset are generated and on the evaluation function used to evaluate the subset under examination [10]. There are three types of procedures for feature subsets generation: Complete, heuristic and random [20]. The evaluation function may be a consistency
measure, a distance measure, an information gain measure or the misclassification error.

Let \( J \) be a feature evaluation function. Assume that a higher value of \( J \) indicates a better feature subset. The function \( J \) has the monotonic property if given two features subsets \( X_1 \) and \( X_2 \) if \( X_1 \subseteq X_2 \), then \( J(X_1) < J(X_2) \), thus, the performance of a feature subset should improve whenever a feature is added to it. Many evaluation functions do not satisfy this monotonic property.

Depending on the generation procedure and the evaluation function [10], the feature selection methods could be divided in two types: filters methods and wrappers methods. A brief description of them is given below.

2.2.1 FILTER METHODS

These methods do not require the use of a classifier to select the best subset of features. Instead they use general characteristics of the data to evaluate features. Among the most important filter methods are: the RELIEF, Las Vegas Filter (LVF) and FINCO, a procedure introduced by Acuña [1]. All these methods are cheap computationally and preserve only the necessary information to perform Knowledge Discovery techniques.

2.2.2 WRAPPER METHODS

These methods use the misclassification error rate of a given classifier as the evaluation function. Instead they use linear discriminant analysis (LDA), k-nearest neighbor (k-nn), and a decision tree classifier called Rpart (it stands for recursive partitioning), which were described in section 2.1.

There are three main approaches to wrapper methods: sequential forward selection (SFS), Sequential Backward selection (SBS), and the sequential floating forward selection (SFFS). A brief description of them is given next:

Sequential forward selection (SFS). This method selects the best feature and then adds the next best feature one such that in combination with the previously selected features maximizes the criterion function. Once that a feature is selected, it cannot be discarded in a later step. It is computationally attractive since to select a subset of size two, it examines only \((m - 1)\) possible subsets.

Sequential Backward Selection (SBS). The principal idea is to see if the classifier can maintain its accuracy by removing one feature at a time until there is only one feature or until accuracy deteriorates to an intolerable level. This method considers initially all the set of variables and then discards the worst feature based on the loss of classification accuracy. Once that a feature is deleted, it can not be taken into account at a later step. A particular case of SBS is Recursive Feature Elimination (RFE) introduced by Guyon and Elisseeff (2003). It uses as underlying classifier, the SVM (Support Vector machine). At each stage the feature with the smallest squared coefficient in the SVM model is deleted. RFE can be generalized by eliminating more than one feature at each stage.

Sequential Forward Floating Search (SFFS). This method is a generalization of the plus-l and away-r method, where first the feature subset is enlarged by \( l \) features using forward selection and then \( r \) features are deleted from the new subset using backward selection. In SFFS the values of \( l \) and \( r \) are determined automatically and update dynamically. This method provides a close optimal subset with an affordable computationally cost.

In this thesis rough set theory will be applied to find an optimal subset of features as suggested in [16, 26].

2.3 Equivalence Relation

Let \( U \) be a non empty set and let \( x, y, \) and \( z \) be elements of \( U \). Consider \( R \) such that \( xRy \) if and only if \( (x,y) \) is in \( R \). \( R \) is an equivalence relation if satisfies:
Reflexive Property: \((x, x)\) is in \(R\) for all \(x\) in \(U\).
Symmetric Property: if \((x, y)\) is in \(R\), then \((y, x)\) is in \(R\).
Transitive Property: if \((x, y)\) and \((y, z)\) are in \(R\), then \((x, z)\) is in \(R\).

2.4 Partition of a set
A partition \(P\) of \(U\) is a family of nonempty subsets of \(U\) such that each element of \(U\) is contained in exactly one element of \(P\).

i) \(U = \bigcup_{i=1}^{n} U_i\),
ii) \(U_i \cap U_j = \emptyset, \ i \neq j\)

![Fig 1. Example of a Partition of a universe set \(U\).](image)

3. ROUGH SETS.
Rough sets theory was proposed by Z. Pawlak (1982). It provides a mathematical tool that can be used to find out all possible feature subsets [19]. The principal idea is to recognize the dispensable and indispensable features, using the discernibility matrix [17, 24]. The purpose of using Rough sets is to find the Core, that is, the set of all indispensable features. The Core does not consider the dispensable features and it can be expanded using Reducts, to finally obtain a good feature subset that permits us to make information induction.

3.1 Indiscernability relation
Rough Sets theory is based on the Indiscernability relation. Let \(T = (U, A, C, D)\), where \(U\) is the universe, \(A\) is a set of features, \(C\) and \(D\) are subsets of \(A\), named the conditional and decision features subsets respectively.

Let \(a \in A, \ P \subseteq A\), the indiscernability relation \(IND(P)\), is defined as follows:

\[
IND(P) = \{(x, y) \in U \times U : \forall a \in P, a(x) = a(y)\}
\]

In simple words, two object are indiscernible if we can not discern between them, because they do not differ enough.
The indiscernability relation defines a partition in \(U\). Let \(U/IND(P)\) denotes a family of all equivalence classes of the relation \(IND(P)\). Two other equivalence classes \(U/IND(C)\) and \(U/IND(D)\), called condition and decision classes respectively, can also be defined.

3.2 Lower Approximation of a subset. Let \(R \subseteq C\) and \(X \subseteq U\), the \(R\)-lower approximation set of \(X\), is the set of all elements of \(U\) which can be with certainty classified as elements of \(X\).

\[
RX = \bigcup\{Y \in U/R : Y \subseteq X\}, \ RX \subseteq X
\]

3.3 Upper Approximation of a subset. the \(R\)-upper approximation set of \(X\) is the set of all element of \(U\), which can be possibly belong to the subset of interest.

\[
\overline{RX} = \bigcup\{Y \in U/R : Y \cap X \neq \emptyset\}, \ X \subseteq \overline{RX}
\]

Any subset defined through its lower and upper approximations is called a rough set, otherwise is called a crisp set.

3.4 Boundary Region.
It is the collection of elementary sets defined by

\[
BNDX = \overline{R}_c X - R_c X
\]

these sets are part of \(R\)-Upper but not of \(R\)-Lower approximations.

3.5 Positive region of a subset. It is the set of all objects from the universe \(U\) which can
be classified which certainty to classes of
U/D employing attributes from C.

\[ \text{POS}_c(D) = \bigcup_{x \in U/D} CX \]

where \(CX\) denotes the lower approximation
of the set \(X\) with respect to \(C\). The positive region of the subset \(X\) belonging to the partition \(U/D\) is also called
the lower approximation of the set \(X\). The union of the positive and the boundary
regions constitutes the upper approximation. The upper approximation contains all data which can possibly be
classified as belonging to the set \(X\) (see Fig. 1).

3.6 Negative region of a subset.
The negative region are those set of
elementary sets that have no predictive
power for a given subset \(X\) given concept,
since they consist of all classes that have no
overlap with the concept. Thus is:

\[ \text{NEG}_A X = (U) - RX \]

3.7 Discernibility Matrix
Let \(U = \{x_1, x_2, x_3, \ldots x_n\}\) be the universe on
a decision table. The Discernibility matrix is defined by:

\[ m_{ij} = \{ a \in C : (a(x_i) \neq a(x_j)) \land (d(x_i) \neq d(x_j), d \in D) \} \]

for \(i, j = 1,2,3,\ldots,n\)

where, \(m_{ij}\) is the set of all attributes that
classify objects \(x_i\) and \(x_j\) into different
decision classes in \(U/D\) partition.

3.8 Dispensable and Indispensable
Features
A dataset contains conditional and decision features [12, 26], some features are
indispensable and are very important in the
analysis. The problem of feature selection is searching indispensable features.
Let \(c \in C\), \(C\) is the set of conditional features. A feature \(c\) is dispensable in the
information dataset \(T\) if \(\text{POS}_{(C-\{c\})}(D) = \text{POS}_c(D):\) otherwise feature \(c\) is
indispensable in \(T\) and should be considered in the final set, the purpose is to retain all
indispensable features that cause \(T\) to be consistent [26]. Thus, if \(c\) is an
indispensable feature, deleting it from \(T\) will cause \(T\) to be inconsistent. In other way if a
feature is dispensable, it could be eliminated from dataset and thus the size is reduced [16].

3.9 Reduct
A system \(T = (U, A, C, D)\) is independent
if all \(c\) in \(C\) are dispensable. A set of
features \(R \subseteq C\) is called reduct of \(C\) if
\(T' = (U, A, R, D)\) is independent and
\(\text{POS}_R(D) = \text{POS}_C(D)\).
A Reduct is a minimal set of features that
preserves the indiscernibility relation
produced by a partition of \(C\). There could be
several subsets of attributes like \(R\). Similar
or indiscernible objects may be represented
several times on an information table, some
of the attributes may be superfluous or irrelevent, and they can be removed without
loss of classification performance.

3.10 The Core
The set of all the features indispensable in \(C\)
in denoted by \(\text{CORE}(C)\). We have

\[ \text{CORE}(C) = \bigcap \text{RED}(C) \]

Where \(\text{RED}(C)\) is the set of all reducts of
\(C\).

The Core is the set of all single element
entries of the discernibility matrix, that is,

\[ \text{CORE}(C) = \{ a \in C : m_{ij} = \{a\} \text{ for some } i, j \} \]

3.11 Dependency coefficient
Let \(T = (U, A, C, D)\) be a decision table.
The Dependency Coefficient between
conditions attributes \(C\), and a decision
attribute \(D\) is defined by the formula:

\[ \gamma(A,D) = \frac{\text{card}(\text{POS}(C,D))}{\text{card}(U)} \]

where \(\text{card}\) indicate cardinality.
3.12 Accuracy of the approximation

The accuracy of the approximation to \( X \) from the elementary subsets is measured as the ratio of the lower and the upper approximation size. The ratio is equal to 1, if no boundary region exists, which indicates a perfect classification. In this case deterministic rules for the data classification can be generated.

3.13 Dependency relation matrix

\[
D(a_i, a_j \mid C) = \sum_{a_i, a_j, y_c} \frac{|\text{Pos}_{a_i}^{y_c}(a_j)|}{\text{card}(y_c)}
\]

\( \text{Pos}_{a_i}^{y_c}(a_j) \) represents the positive region of attribute \( a_j \) relative to attribute \( a_i \) within the class value \( y_c \) [19].

3.14 Rule discovery based on rough set theory.

Rule discovery is an important problem since data relationships in the form of "if \( A \) then \( B \)" do not necessarily reflect real rules of the application domain, also other problems could arise. Therefore, there is a need to eliminate incorrect rules. The construction of a rule discovery algorithm with associated estimates of error rates of classification could be developed by the rough sets theory [23].

4. DATA PREPROCESSING USING ROUGH SETS

In this section potential areas of research to be considered in this thesis are discussed, Rough sets is used to build algorithms to calculate reducts in order to perform feature selection. Also, algorithms for discretization and imputation of missing values based on rough sets will be developed. These methods will be combined with: lda, k-nn, and rpart classifiers to evaluate their effect on the classification error. Instance selection procedures will be other aspect to be considered.

4.1 FEATURE SELECTION

Similarity relations can be used to evaluate a subset of feature. Thus, rough sets theory offers us a new alternative to select the dispensable features subset.

On the other hand, in this thesis, evaluation functions for feature selection based on Rough set theory will be proved to be monotonic.

4.2 INCOMPLETE DATA.

When the dataset has missing attribute values, there are two options: deleted the instances containing them or apply a method to impute them. A principal assumption to analyze incomplete data with rough set is that it missing values are lost, then the attribute value blocks are used by rough set theory to construct characteristic sets indiscernability relation and lower and upper approximations [8].

Rough sets is one of the techniques for identification and recognition of common patterns in data, especially in the case of uncertain and incomplete data. The purpose of doing imputation of missing values based on rough set theory is to use the available data to replace the missing values using the indiscernability relation.

4.3 CASE SELECTION

Instance selection in a dataset is carried out to obtain an appropriate subset of instances to perform a KDD task [21]. Each group of features has some instances like representative of elementary blocks, then extracting a subset of instances is related to set weights to each elementary sets produced by Rough sets theory, and in this way extract interesting instances. Some instances in the dataset are inconsistent because they might have all their feature values similar to other instances but lie in a different class. These instances should be analyzed with care. Elementary sets formed using the set \( C \) of conditional features help to identify the weight class where there should be inconsistent instances. Instances selection reduces the computation time of executing some KDD task.
4.4 DISCRETIZATION.
Rough set theory is based on decision tables. According to this, we need to discretize continuous features. Some discretization methods commonly used are based on: entropy and the chi-square statistic. There are methods based on equal width intervals or equal number of instances in each interval [11,15]. Discretization methods that use the class label, are referred to as supervised discretization methods [6,22]. In this thesis, discretization methods based on rough set theory to find the cut points will be introduced. Using the similarity measure based on rough set, some variants of the standard methods will be introduced.

5. UNSUPERVISED LEARNING BASED ON ROUGH SETS
Unsupervised classification or clustering describes the subdivision of the universal set of all possible categories into a number of distinguishable categories called elementary sets. Hierarchical and non-hierarchical clustering methods use a membership function and a similarity measure [5]. Each elementary set can be regarded as a rule describing the object of the classification [26]. Then, each object is classified using the elementary set of features which can not be split up any further, although other elementary sets of features may exist. In this thesis application of rough sets to clustering will be discussed, and new methods will be proposed.

6. APPLICATIONS
Bioinformatics is an interesting area where many knowledge discovery tasks can be applied. Among them supervised or unsupervised classification (clustering). All methods developed using Rough set theory could be applied to analyze bioinformatics data. The number of features in a bioinformatics’ dataset is large, and there could be a large amount of missing values. Rough set theory offers methods to handle the problem of missing values [12]. The datasets to be considered in this thesis will be the same used in similar studies, including some gene expression data from microarrays experiments.

7. ETHICS
Rough set is a mathematical tool to analyze uncertain information systems. Therefore, much professionalism is required when handling data. Data gathering is very important, thus one must try not to significantly manipulate data, because such practice leads to inappropriate results.

The researcher has responsibilities about his analysis, interpretation, and publications [3]. Hence, it is necessary to write with consideration and to be careful on the research process until the task is finished and published.

Ethics is a subject studied in the literature of philosophy specially the philosophy of information technologies, like Kant's categorical imperative, having presented his view on how and why something may be considered moral: "Act only according to that maxim by which you can at the same time will that it would become a universal law." [10]. When the world (the object) is viewed or analyzed we know a priori that morality is universal and necessary then, applied science according to ethical rules is a moral thing.

Rough sets it’s a mathematical tools that helps development of the Knowledge discovery methods. Given the diversity of these methods applications, they can be given inappropriate use.
Bibliography


IEEE Transactions on Neural Networks. pp. 3 - 14


