ÉVA KOVÁCS

Contributions to the Development of Methods in Data Mining in Large Databases

PHD THESIS

Abstract

Scientific advisor
Prof. Dr. Ing. IOSIF IGNAT

__________________ Cluj-Napoca, 2007__________________
## Contents

### Chapter 1. Introduction
- 1.1. Definitions
- 1.2. Description of the process of data mining
- 1.3. Preprocessing and transformation of data for data mining
- 1.4. Operations in data mining
- 1.5. Clustering methods of the database
- 1.6. Predictive modeling methods
- 1.7. Link analysis methods
- 1.8. Structure of the thesis

### Chapter 2. Current State of Data Mining
- 2.1. Clustering methods
- 2.2. Classifications methods
- 2.3. Feature selection methods
- 2.4. Conclusions

### Chapter 3. Development of Clustering Methods with Clustering with Prototype Entity Selection
- 3.1. Clustering with Prototype Entity Selection
- 3.2. Clustering with Prototype Entity Selection Algorithm
- 3.3. Optimized CPES using variable radius
- 3.4. CPES Algorithm with variable radius
- 3.5. K-means optimized by CPES
- 3.6. Experimental study
- 3.7. Results of the experimental study for CPES
- 3.8. Results of the experimental study for CPES with variable radius
- 3.9. Results of the experimental study for K-means with CPES
- 3.10. Analysis of the results
- 3.11. Conclusions

### Chapter 4. Development of Preprocessing Methods by Reduct Equivalent Feature Selection
- 4.1. Rough Set Theory
- 4.2. Reduct Equivalent Feature Selection
- 4.3. Reduct Equivalent Feature Selection Algorithm
- 4.4. Experimental Study
- 4.5. Conclusions

### Chapter 5. Development of Classification Methods by Reduct Equivalent Rule Induction
- 5.1. Classification from the perspective of Rough Set Theory
- 5.2. Reduct Equivalent Rule Induction
- 5.3. Reduct Equivalent Rule Induction Algorithm
- 5.4. Experimental Study
- 5.5. Conclusions

### Chapter 6. Final Conclusions
- 6.1. Future research trends

### Bibliography

### Annexes
Chapter 1.
Introduction

The development of hardware during the last two decades made possible the saving of large amounts of data on computers. The exact volume of these data is impossible to state, all estimations are mere suppositions. Researchers from the Berkley University have calculated that approximately one Exabyte (1 million Terabytes) of data are generated and saved each year.

Real-time data mining of databases is one of the main research areas in databases. The volume of databases and their constant growth is an important problem in data mining. Data mining is a new discipline in development which uses the resources and ideas of several fields. The abstract role of data mining is to discover new and useful information in databases. Data mining techniques are to develop models, structures, regularities, etc, in large databases. The models discovered in databases can be characterized according to: accuracy, precision, interpretability and expressivity.

Knowledge can be defined in several ways, for example:

- General term used to describe an object, idea, condition, situation or another fact that can be a number, a letter or a symbol. It can be a chart, an image and/or alphanumeric characters. It suggests elements of information that can be processed or produced by a computer.

- Facts, known things, one can draw conclusions from.

There are several definitions for data mining; we mention only few of them:

- “Data mining consists of the extraction of predictive information hidden in a large database.”
- “Data mining is the process through which advantageous models are discovered in databases”
- “Data mining is a rapidly growing field that combines the methods of databases, statistics, supervised learning and other related fields in order to extract useful information from existing data.”
- “Data mining is the non-trivial process which identifies new, valid, useful and interpretable models in existing data.”

Algorithms of data mining can be categorized keeping in mind the representation of models, input data and the field in which the algorithm is used. The model can be represented by decision trees, regression functions, associations or others. The most common classification of data mining algorithms divides them in four operations: predictive modeling, database clustering, link analysis and deviation detection [6][7][8][9].

Data mining is a repetitive process, which contains several steps, beginning with the understanding and definition of the problem and ending with the analysis of the results and the application of a strategy for the use of the results [14]. A data mining process is illustrated in Fig. 1.

![Fig. 1. The data mining process](image)
Chapter 2.
Current State of Data Mining

The second chapter describes the present state of research on data mining, and the most recent research in the field. The algorithms of the main methods in data mining are then presented, namely that of database clustering, predictive modeling and feature selection.

From the analyzed methods used in database clustering, from the category of partitioning methods K-means, K-modes, K-medoid, CLARA (Clustering Large Applications) and CLARANS (Clustering Large Application based upon RANdomized Search) are mentioned. K-means algorithm is one of the best known clustering algorithms used in database clustering. The necessity that the user should specify the number of clusters is a disadvantage. The method is also inadequate for finding clusters of nonconvex forms or of different sizes and is sensitive to noises that influence the mean value of a cluster. K-medoid algorithm is also a clustering algorithm used in database clustering. It works efficiently with a small set of data, while it is unable to manage a large set. CLARA (Clustering LARge Applications) is used instead of K-medoid in order to manage a large set of data.

From the analyzed methods used in database clustering SLINK (Single LINKage) and BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) belong to the category of hierarchical methods, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and OPTICS (Ordering Points To Identify the Clustering Structure) to the category of density-based methods.

From the analyzed methods within predictive modeling from the category of algorithms that are based on decision trees we mention ID3, C4.5, CHAID (Chi-square automatic interaction detection), CART (Classification And Regression Trees) and QUEST (Quick, Unbiased and Efficient Statistical Tree).

The most popular classification algorithm used in predictive modeling on neural networks is backward propagation algorithm. It learns on a multilayer feed-forward network. Neural networks are more efficient than decision trees due to the adjustment. A deficiency of neural networks is that they accept only numeric input, therefore categorical data must be recoded.

Bayesian classifiers are statistical classifiers and belong to predictive modeling. Bayesian classifiers demonstrated high precision and speed, and are used for large databases. In practice several disagreements appear, for example, due to the supposition that attributes are independent from one another.

The majority of feature selection methods belong to supervised learning. There are two types of feature selection algorithms: filter and wrapper type algorithms. RELIEF algorithm and its developments are representative for filter type. There are feature selection methods that use successfully Rough Set Theory.

Chapter 3.
Development of Clustering Methods with Clustering with Prototype Entity Selection

This chapter presents three original clustering methods: Clustering with Prototype Entity Selection (or CPES) [16][19], Clustering with Prototype Entity Selection with variable radius [17] and K-means with Clustering with Prototype Entity Selection [18]. First the methods, then the experimental studies are presented, followed by the comparison of the results with other clustering methods from the specialized literature.

CPES method will be used in data mining both as an independent clustering process and a combination of this method with K-means, a clustering algorithm often used in data mining.
Clustering with Prototype Entity Selection Algorithm

We propose the use of CPES method as a clustering method for data mining. When using this method the user does not need to specify the number of clusters, the algorithm will obtain this number. The method also ensures optimal clustering. This method will be efficient within the frame of a data mining process, because it is very important for a user that the used method should need only few input data. In short, the algorithm is presented as follows:

1. Initialization of constants $d_{av}$, $A$, $r$ and fitness function $f$
2. Generation of clusters $\text{cluster}(x_i) = x_i$, for $i=1,...,n$
3. repeat
   3.1. For each $\text{cluster}(x_i)$ a pair is selected, $x_j$
   3.2. if $f(\text{cluster}(x_i)) < f(x_j)$ then
       set $\text{cluster}(x_i) = x_j$
   until there are no more changes in the clusters

Then the steps of the algorithm are described:

1. Constants $d_{av}$, $A$, $r$ are initialized and the values of fitness function $f$ are calculated for each object of the data set, using:
   
   \[
   f(x_i) = \frac{1}{\sum_{k=1}^{n} d(x_i, x_k) + A} \quad (1)
   \]
   
   \[
   d_{av} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} d(x_i, x_j)}{n(n-1)} \quad (2)
   \]
   
   \[
   A = \frac{d_{av}}{n} \quad (3)
   \]
   
   \[
   r = \frac{d_{av}}{2} \quad (4)
   \]

2. After initialization $n$ clusters are defined, each object will have its own cluster, namely $\text{cluster}(x_i) = x_i$ for each $i=1,...,n$. $\text{cluster}(x_i)$ is the cluster $x_i$ belongs to. In this step the value of cluster $\text{cluster}(x_i)$ is just $x_i$ and there are $n$ clusters with different values.

3. The cycle is repeated until stop condition is carried out, i.e. until there are no more changes in the values of $\text{cluster}(x_i)$ for each $i=1,...,n$ from one step to the other.

3.1. For each $\text{cluster}(x_i)$ a pair is selected. This pair is chosen from the objects of $\text{cluster}(x_i)$ so that $i \neq j$ and the condition that $\text{cluster}(x_j)$ is in the radius $r$ of object $\text{cluster}(x_i)$, i.e.

\[
\text{cluster}(x_j) \in V(\text{cluster}(x_i), r) \quad (5)
\]
For the selection, proportional selection is used and the chosen pair is marked with \( \text{cluster}(x_p) \).

3.2. The values of function \( f \) are compared for the objects of \( \text{cluster}(x_i) \) and \( \text{cluster}(x_p) \), and the object that has the highest value of fitness function \( f \) is chosen. If the condition \( f(\text{cluster}(x_i)) < f(\text{cluster}(x_p)) \) is true, then the values of cluster \( x_i \) are set with the new value \( \text{cluster}(x_p) \). If the condition is not true, then the value of cluster \( x_i \) remains unchanged.

In each step of the algorithm, the number of the different clusters will diminish. In the end there will be only \( m \) distinct values in \( \text{cluster}(x_i) \). These are marked with \( c_j, j=1,\ldots,m \). These distinct values can be considered prototype entities for their cluster, and objects \( x_i \) for which \( \text{cluster}(x_i) = c_j \) belong to cluster \( c_j \). The number of clusters thus obtained will be equal to \( m \) and it is determined for each object the cluster it belongs to.

By CPES method a clean clustering is obtained with an optimal number of clusters for input data. If the analyst is not satisfied with the accuracy of the clusters, another algorithm of clustering can be performed starting with the results already obtained by the use of the CPES method.

**Optimized CPES Using Variable Radius**

The improvement of the original clustering method, *Clustering with Prototype Entity Selection*, is presented. The radius with variable value is used to optimize the CPES method.

The CPES method starts with an initialization phase in which radius \( r \) is calculated. During the CPES algorithm, radius \( r \) remains constant and the partner of object \( x_i \) will be chosen from the vicinity of \( V(x_i,r) \) of \( x_i \). As the pair of object \( x_i \) is chosen from the vicinity of \( V(x_i,r) \), there will be no massive migration among clusters. Still there is a migration among clusters, therefore the CPES algorithm clusters badly some of the objects.

In order to optimize CPES algorithm instead of radius \( r \), we use a radius with variable value. Thus at the beginning radius \( r \) is set with the average distance divided by 4, \( r = r_1 = \frac{d_{av}}{4} \). After each step of the algorithm that is performed, the value of the radius increases, so the migration among clusters will be smaller than in the case of a radius with fixed value.

A diagram is used to increase the value of the radius, defined by the following equation:

\[
 r_{k+1} = r_k + \alpha \frac{d_{av}}{4}
\]

where \( \alpha \) is a positive constant smaller than 1.

**K-means Optimized by CPES**

The improvement of the *K-means* algorithm is shown by using *Clustering with Prototype Entity Selection*, (CPES); thus the user does not need to introduce the number of optimal clusters. This method was developed to be used in data mining. It presents the advantages of the CPES algorithm in boosting the performances of *K-means* method and the experimental results obtained as compared to *K-means* method.
By using CPES method we find out both the optimal number of clusters in the data set proposed for data mining, \(m\), and the prototype entities, \(c_j, j=1,...,m\). These data will be used as input data for K-means.

K-means method will not start with the arbitrary selection of a representative object for each cluster; instead it will use the prototype entities obtained due to CPES algorithm, \(c_j, j=1,...,m\). These prototype entities, \(c_j\) represent the mean, the centre of the clusters. As they are local maximums, there will be no massive migration among the clusters formed during the first step. K-means algorithm will need a short running time. Because the number of clusters is generated by the CPES method, and not specified by the user, an optimal number of clusters will be generated. In what follows K-means method will be carried out as usual, namely each object is attributed to that cluster to which it resembles the most. The mean of each cluster is recalculated, then each object is reattributed to the clusters, having in mind the new mean of the clusters.

Consequently K-means algorithm will be improved by CPES, due to the followings:

- The user does not need to state the number of clusters, as they will be generated by CPES.
- The algorithm does not need to choose accidentally the first representative objects, they will be prototype entities generated by CPES.

The described optimization was presented within the framework of K-means method. Nevertheless this optimization can be also applied to other clustering methods which need as input data the number of clusters, for example for K-medoid.

**Chapter 4.**

**Development of preprocessing methods by Reduct Equivalent Feature Selection**

A major problem in data mining is the large dimension of data. Therefore it is very important that during the preprocessing of data for data mining, irrelevant attributes to be eliminated without modifying the consistency of the original set of data. Feature selection is the process by which an optimal subset of attributes that satisfy certain criteria is searched for. This chapter presents Reduct Equivalent Feature Selection [20], (REFS) a new heuristic method of feature selection based on Rough Set Theory. It describes mathematically the essence of Reduct Equivalent Feature Selection method, and presents the algorithm and the obtained experimental results.

**Reduct Equivalent Feature Selection**

*Rough Set Theory* is successfully used in data mining, or more precisely in prediction, in discovery of associations and in reduction of attributes. Having as motto “Let data speak for themselves”, this approach is not invasive for the processed data set as it uses only the information of the data set without presuming the existence of different models in it [28][30].

The reduction of attributes using *Rough Set Theory* means the finding of a reduct set from the original set of attributes. Data mining attributes will not run on the original set of attributes but on this reduct set, which will be equivalent with the original set.

Two original theorems were devised: Theorem 1 and Theorem 2. They were developed to find a reduct quickly. Theorem 2 was used to find a set of reduct for a data set. The algorithm that will use this theorem is called Reduct Equivalent Feature Selection. It is a new heuristic method of feature selection that will be presented in the following.
Reduct Equivalent Feature Selection Algorithm

Reduct Equivalent Feature Selection is a new heuristic method of feature selection based on Rough Set Theory. A forward selection will be used that begins with an empty set of attributes to which in each step of the algorithm a new attribute is added until a set of reduct attributes is obtained.

Reduct Equivalent Feature Selection generates in each step a new set of attributes and evaluates it, i.e. verifies if the generated set is or is not a reduct set. If the obtained set is not a reduct set, the generation is resumed and another element is added to the proposed set of attributes. This new set of attributes is evaluated again. If the set of attributes proposed is a reduct set, then it is validated and can be used in the future too as a set equivalent with the original set of attributes.

Input data of the algorithm will be data set \( U = U_0 \), the attributes of condition \( C = R_0 \) and those of decision \( D \). The method results in a reduct set, \( R \) of the attributes of condition \( C \) taking into consideration the attributes of decision \( D \).

In the first step of the algorithm reduct set, \( R \) is initialized with an empty set of attributes and variable \( i \) with number 0.

The second step is repeated until \( POS_R(D) = POS_C(D) \), i.e. until the set \( R \) of attributes will be a reduct. In this step the core set of the set of attributes \( R_i \) is calculated. If the core of set \( R_i \) is an empty set, it means that any other attribute \( a \) can be removed from \( R_i \) without modifying the relation of indiscernibility. There is no attribute \( a \) in \( R_i \) that would be indispensable. \( POS_C(D) = POS_{C \setminus \{a\}}(D) \), for any \( a \in R_i \). Thus an attribute is removed from \( R_i \) and the core set for the new \( R_i \) is searched until the core set is no more an empty set. This core set is added to set \( R \). Then number \( i \) is increased with 1 and the new \( R_i \) and \( U_i \) is calculated for the next step of the cycle.
Chapter 5.
Development of Classification Methods by Reduct Equivalent Rule Induction

Rough Set Theory is successfully used in data mining for prediction, classification \([1][3][22][23][29]\), discovery of associations \([11]\) and reduction of attributes \([27][32]\). This chapter presents original contributions to the classifications of the objects of a database using the elements of Rough Set Theory. A new classification method based on Rough Set Theory is presented: Reduct Equivalent Rule Induction (or RERI). This chapter describes the mathematical essence of Reduct Equivalent Rule Induction method, as well as the algorithm and the obtained experimental results.

This chapter presents the way in which Rough Set Theory was used to develop a new classification method. Defined reduct sets in Rough Set Theory produce minimal rules of classification among the attributes of the analyzed data set. Such a rule of classification determined by a reduct contains a minimal number of attributes in its condition part, so that the set of rules of classification defines the classes of decision.

Reduct Equivalent Rule Induction

We look for classification rules for a data set \(U\). If all the attributes of the set are taken into consideration in order to generate rules of classification, too many rules of classification will appear. They are also going to be too detailed and incapable of predicting to which class a new entry belongs to. If too few attributes of the data set are taken into consideration in order to characterize the whole data set through rules of classification, the rules thus obtained will be too general. They will be also unable to predict correctly to which class a new entry belongs to.

We propose the use of a reduct set to generate rules of classification. These rules can be then used to classify a new record. As a reduct is a minimal set of attributes that maintains the positive region of classes of decision \(U/\text{IND}(D)\) taking into consideration the attributes of condition in \(C\), the quality of the rules of classification obtained remains uncompromised. By reducing the number of the used attributes in rules of condition a precise classification is ensured, because only the redundant or irrelevant attributes of classification are eliminated.

It describes a new, original and efficiently used method to classify the data of a data set \(U\), namely Reduct Equivalent Rule Induction. This method proposes the use of a reduct to generate rules of classification. These rules of classification will be used to classify new entries. The method proposed uses the Reduct Equivalent Feature Selection algorithm described in chapter 4 to generate the reduct set. The obtained reduct set will maintain the positive region of the classes of decision \(U/\text{IND}(D)\) taking into consideration the attributes of condition in \(C\). From the reduct set Reduct Equivalent Rule Induction method generates the minimal rules of classification that will be used to classify a new entry.
Chapter 6.
Final Conclusions

The general objective of this thesis was the development of data mining methods in large databases. Data mining techniques are used to discover models, structures, regularities, etc, in very large databases. In order to accomplish the objectives of this thesis new and original methods have been developed that are useful for data mining techniques.

Chapter 2 describes the present state of research in data mining. This chapter presents the main data mining methods, as well as the most recent research in this field. It also presents the algorithms of main methods in data mining, namely that of database clustering, predictive modeling and link analysis.

Chapter 3 presents original contributions to the development of data mining methods: Three new clustering methods are present: Clustering with Prototype Entity Selection (CPES), Clustering with Prototype Entity Selection with variable radius and K-means with Clustering with Prototype Entity Selection. These methods were developed to be used in data clustering.

The main advantage of the methods proposed in chapter 3 is that they do not need the number of clusters as input data. The algorithms find by themselves the optimal number of clusters, then cluster all the objects in these clusters. The methods proposed in this chapter can be used easily by the analyst and have better or as good results as those methods that need complicated settings of the used parameters.

Chapter 3 contains the following personal contributions:
1. I proposed the use of certain elements of genetic algorithms, such as fitness function and proportional selection to create a new clustering algorithm for data mining.
2. I proposed a new algorithm Clustering with Prototype Entity Selection to be used in data clustering, an algorithm that does not need as input data the number of clusters.
3. I proposed the use of a variable radius to optimize Clustering with Prototype Entity Selection, formulating the new algorithm as follows: Clustering with Prototype Entity Selection with variable radius.
4. I optimized K-means method by using Clustering with Prototype Entity Selection algorithm, formulating thus the new algorithm, K-means with Clustering with Prototype Entity Selection. In order to optimize K-means method I used the number of clusters and prototype entities generated by CPES.
5. The presented optimization in the context of K-means method can be applied to other methods of clustering that need as input data the number of clusters, for example in K-medoid method.
6. I implemented the three proposed methods: Clustering with Prototype Entity Selection, Clustering with Prototype Entity Selection with variable radius and K-means with Clustering with Prototype Entity Selection.
7. I compared the results of the proposed methods in chapter 3 with other clustering algorithms used in data mining, respectively with K-means and SLINK algorithms.

Chapter 4 presents original contributions to dimensionality reduction in preprocessing of data for data mining. It describes Reduct Equivalent Feature Selection, a new, original, heuristic method of feature selection based on Rough Set Theory. This method eliminates irrelevant attributes without modifying the consistency of the original data set.

Chapter 4 contains the following personal contributions:
1. I devised a new theorem that offers an original calculation method of a reduct set, Theorem 1.
2. Generalizing Theorem 1, I devised Theorem 2, this is also a new method for calculating a reduct set.
3. Using Theorem 2, I proposed a new algorithm: Reduct Equivalent Feature Selection to be used in feature selection.
4. I implemented the proposed method, Reduct Equivalent Feature Selection.
5. I tested the proposed method in Chapter 4 with 10 sets of different data. These data sets were created and used by several researchers and are reference examples for data mining and belong to the UCI collection.
Chapter 5 presents original contributions to the classification of the objects of a database using elements of Rough Set Theory: A new classification method is presented, Reduct Equivalent Rule Induction (or RERI) based on Rough Set Theory.

Chapter 5 contains the following personal contributions:

1. Using Theorem 2 described in Chapter 4, I proposed a new algorithm of classification, Reduct Equivalent Rule Induction.
2. I implemented the proposed method, Reduct Equivalent Rule Induction.
3. I tested the proposed method in Chapter 5 with 10 sets of different data. These data sets were created and used by several researchers and are reference examples for classification and data mining and belong to the UCI collection.
4. I compared the accuracy of the new algorithm RERI with other classification methods from specialized literature, CHAID and QUEST.

Data mining is a field that can be developed in many directions. Data mining process should be as automatic as possible, without the excessive intervention of the analyst. In order to achieve the highest possible automatism of the system, it must be capable of preprocessing correctly and optimally the data, so that data mining algorithms can extract useful information from it.

Due to the study accomplished, based both on specialized literature and the results obtained in this thesis, I propose some new research directions:

- Optimization of clustering methods that need as input data the number of clusters.
- Use of Rough Set Theory in data mining within the framework of discovering of associations.
- Use of Rough Set Theory in reducing the dimension of the data set proposed for data mining.

Selected references


