IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

THEORY APPROACH ON ROUGH SET

A.A.Narasimham*
Department of Computer Science and Engineering, DIET college, Anakapalle, Visakhapatnam Andhra Pradesh, India.

ABSTRACT

Unsupervised clustering is an essential technique in Data mining. Rule identification involves the application of Data mining techniques to derive usage patterns from the information system. Knowledge extraction from data is the key to success in many fields. Knowledge extraction techniques and tools can assist humans in analyzing mountains of data and to turn the information contained in the data into successful decision making. This paper proposes, to consider an information system without any decision attribute. The proposal is useful when we get data, which contains only input information (condition attributes) but without decision (class attribute). K-Means algorithm is applied to cluster the given information system for different values of K. Decision table could be formulated using this clustered data as the decision variable.

KEYWORDS: Data mining-Means Clustering Rough set,.

INTRODUCTION

Data mining refers to extracting or "mining" knowledge from large amounts of data. There are many other terms carrying a similar or slightly different meaning to Data mining, such as knowledge mining from databases, knowledge extraction, data pattern analysis, data archaeology, and data dredging. Data mining treats as synonym for another popularly used term, Knowledge Discovery in Databases (KDD) [9]. KDD consists of the following steps to process it such as Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation and Knowledge presentation.

KDD is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Data mining is not a single technique, some commonly used techniques are: Statistical Methods, Case-Based Reasoning (CBR), Neural Networks, Decision Trees, Rule Induction, Bayesian Belief Networks (BBN), Genetic Algorithms, Fuzzy Sets and Rough Sets. Data mining relates to other areas, including machine learning, cluster analysis, regression analysis, and neural networks. Both neural network and regression approaches create the same model based on a training data set. This model normally uses a predetermined set of features. A machine learning algorithm of data mining generates a number of models (usually in the form of decision rules) capturing relationships between the input features and the decisions. In an extreme case, the set of features included in each rule could be independent from all other rules, which is similar to the result produced by cluster analysis. Neural network and regression models can be viewed as "population based" as a single model is formed for the entire population (training data set), while the data mining approach follows an "individual (data object) based" paradigm. The "population based" tools determine features that are common to a population (training data set). The models (rules) created by data mining are explicit.

One of the new data mining theories is the rough set theory [28] that can be used for

(i) reduction of data sets
(ii) finding hidden data patterns
(iii) generation of decision rules

M. Goebel et.al [8] provided an overview of common knowledge discovery tasks and approaches to solve these tasks. A novel clustering technique is applied to cluster the information system in order to get class attribute. The overview of the clustering technique is elaborated here under.

Clustering Overview:

Clustering [9] is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Clustering has its roots in many areas, including Data mining, statistics, biology, and machine learning. Clustering is a type of multivariate statistical analysis also known as cluster analysis, unsupervised classification analysis, or numerical taxonomy. Cluster analysis is based on a mathematical formulation of a measure of similarity. In other words cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. In other words, cluster analysis simply discovers structures in data without explaining why they exist.

Clustering is an excellent example of an unsupervised learning technique [6,7] . In general, we can apply the v-fold cross-validation method to a range of numbers of clusters in K-Means clustering, and observe the resulting average distance of the observations (in the cross-validation or testing samples) from their cluster centers (for K-Means clustering). In this paper, the original decision table together with a new decision attribute obtained by Self-Organizing Maps (SOM) is reconstructed. The SOM is applied as a cluster method.

Rough Set Based Feature Reduction:

In 1982, Pawlak introduced the theory of Rough sets [28,29]. This theory was initially developed for a finite universe of discourse in which the knowledge base is a partition, which is obtained by any equivalence relation defined on the universe of discourse. In rough sets theory, the data is organized in a table called decision table. Rows of the decision table correspond to objects, and columns correspond to attributes. In the data set, a class label to indicate the class to which each row belongs. The class label is called as decision attribute, the rest of the attributes are the condition attributes. Here, C is used to denote the condition attributes, D for decision attributes, where $C \cap D = \Phi$, and t_j denotes the j^{th} tuple of the data table. Rough sets theory defines three regions based on the equivalent classes induced by the attribute values: lower approximation, upper approximation, and boundary. Lower approximation contains all the objects, which are classified surely based on the data collected, and Upper approximation contains all the objects, which can be classified probably, while the boundary is the difference between the upper approximation and the lower approximation. Hu et al., [10] presented the formal definitions of rough set theory. A. Kusiak [19] described the basic concepts of rough set theory, and other aspects of Data mining. The other aspects of data mining are Equivalence classes, Atoms, Approximation accuracy, Boundary approximation, Classification accuracy, Classification quality, Sensitivity, Specificity, Positive predicted value, Negative predicted value, Rule length, Rule strength, Exact rule, approximate rule, Rule support, Rule coverage, Rule acceptance and Discrimination level.

Let U be any finite universe of discourse. Let R be any equivalence relation defined on U, which partitions U. Here, (U, R) which is the collection of all equivalence classes, is called the approximation space. Let W1, W2, W3, …, Wn be the elements of the approximation space (U, R). This collection is known as knowledge base. Then for any subset A of U, the lower and upper approximations are defined as follows:

$RA = \cup\{Wi / Wi \subseteq A\}$

$RA = \cup \{Wi / Wi \cap A \neq \emptyset\}$

The ordered pair (RA, RA) is called a rough set. Once defined these approximations of A, the reference universe U is divided into three different regions: the positive region POSR(A), the negative region NEGR(A) and the boundary region BNDR(A), defined as follows:

POSR(A) = RA

NEGR (A) = U – RA

BNDR (A) = RA – RA

Hence, it is trivial that if BND(A) = Φ, then A is exact. This approach provides a mathematical tool that can be used to find out all possible reducts. However, this process is NP-hard [17, 45], if the number of elements of the universe of discourse is large. As there is a one-to-one correspondence between the knowledge base and knowledge representation, the theory can be adopted for the decision tables in information systems.

Feature selection process refers to choose subset of attributes from the set of original attributes. Feature selection has been studied intensively for the past one decade [15,16,21,26]. Besides that the brief introduction given here, the extensive literature of Rough sets theory can be referred to Orlowska [27], Peters and for recent comprehensive and overviews of developments.

The purpose of the feature selection is to identify the significant features, eliminate the irrelevant of dispensable features to the learning task, and build a good learning model such as web categorization discussed in [13]. The benefits of feature selection are twofold: it considerably decreased the computation time of the induction algorithm and increased the accuracy of the resulting mode. All feature selection algorithms fall into two categories: (1) the filter approach and (2) the wrapper approach. In the filter approach, the feature selection is performed as a preprocessing step to induction. The filter approach is ineffective in dealing with the feature redundancy. Some of the algorithms in the Filter approach methods are Relief, Focus, Las Vegas Filter (LVF), Selection Construction Ranking using Attribute Pattern (SCRAP), Entropy-Based Reduction (EBR), Fractal Dimension Reduction (FDR). In Relief [16] each feature is given a relevance weighting that reflects its ability to discern between decision class labels. Focus [1], conducts a breadth-first search of all feature subsets to determine the minimal set of features that can provide a consistent labeling of the training data. LVF employs an alternative generation procedure – that of choosing random features subsets, accomplished by the use of a Las Vegas algorithm [22] is an instance based filter, which determines feature relevance by performing a sequential search within the instance space. EBR [11] based on the entropy heuristic employed by machine learning techniques such as C4.5. EBR is concerned with examining a dataset and determining those attributes that provide the most gain in information is a novel approach to feature selection based on the concept of fractals – the self-similarity exhibited by data on different scales. In the wrapper approach [15], the feature selection is “wrapped around” an induction algorithm, so that the bias of the operators that defined the search and that of the induction algorithm interact mutually. Though the wrapper approach suffers less from feature interaction, nonetheless, its running time would make the wrapper approach infeasible in practice, especially if there are many features, because the wrapper approach keeps running the induction algorithms on different subsets from the entire attributes set until a desirable subset is identified. We intend to keep the algorithm bias as small as possible and would like to find a subset of attributes that can generate good results by applying a suite of data mining algorithms. Some of the Wrapper approach methods are Las Vegas Wrapper (LVW) and Neural network-based feature selection. The LVW algorithm [23] is a wrapper method based on LVF algorithm. This again uses a Las Vegas style of random subset creation which guarantees that given enough time, the optimal solution will be found.

Neural network-based feature selection is employed for backward elimination in the search for optimal subsets. A decision table may have more than one reduct. Anyone of them can be used to replace the original table. Finding all the reduces from a decision table is NP-Hard [20]. Fortunately, in many real applications it is usually not necessary to find all of them and it is enough to compute one such reduct is sufficient [10]. A natural question is which reduct is the best if there exist more than one reduct. The selection depends on the optimality criterion associated with the attributes. If it is possible to assign a cost function to attributes, then the selection can be naturally based on the combined minimum cost criteria. In the absence of an attribute cost function, the only source of information to select the reduct is the contents of the data table [26]. For simplicity, we adopt the criteria that the best reduct is the one with the minimal number of attributes and that if there are two or more reducts with same number of attributes, then the reduct with the least number of combinations of values of its attributes is selected have applied Rough Sets with Heuristics (RSH) and Rough Sets with Boolean Reasoning (RSBR) for attribute selection and discretization of real-valued attributes.

In Section 3 the Data mining techniques are studied and implemented using MATLAB for the various data sets obtained from UCI machine learning repository [5] and the real HIV data set. Section 3 describes the experimental analysis of Quickreduct and VPRS, Section 4 states the conclusion of this paper and the directions for further research are proposed herein.

LITERATURE REVIEW

As mentioned that, data mining techniques can be classified along several dimensions. The most extensive effort in this direction has been done and they have only focused on data mining techniques developed in machine learning and statistical domains. Most of the other reviews on data mining techniques have chosen to focus on a particular sub-area of the existing research. Markou and Singh presented an extensive review of novelty detection techniques using neural network sand statistical approaches. A review of selected data mining techniques, used for network intrusion detection. Data mining techniques developed specifically for system call intrusion detection have been reviewed by. A substantial amount of research on data mining has been done in statistics as well as other reviews provide a unification of several distance based data mining techniques. These related efforts have either provided a coarser classification of research done in this area or have focused on a subset of the gamut of existing techniques. To the extent of our knowledge, our survey is the first attempt to provide a structured and a comprehensive overview of data mining theory.

DATA MINING PROCESS

The block diagram of the data mining methodology is depicted in the following figure.

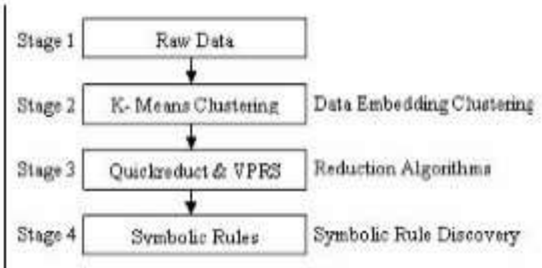

Figure 1. Datamining Process

A) Data Preparation

In the first stage, the data sets viz., Iris, Zoo, and Soybean (small) obtained from UCI machine learning repository [5] and the real HIV data set are considered for this study and it is tabulated in the Table 6. The HIV database consists of information collected from the HIV Patients at Voluntary Counseling and Testing Centre (VCTC) of Government Hospital, Dindigul District, Tamilnadu, India, a well-known centre for diagnosis and treatment of HIV. The advantage of this data set is that it includes a sufficient number of records of different categories of people affected by HIV. The set of descriptors presents all the required information about patients. It contains the records of 500 patients. The record of every patient contains 49 attributes and this has been reduced to 22 attributes after consulting the Physician. The details of attributes are given as follows: The continuous attributes are Age, Sex, Marital-Status, Occupation, Area, Loss-of-Weight, Continuous-Fever, Continuous-Cough, Skin-Disease, Oral-Thrush, Tuberculosis, Diarrahoea, Anaemia, Sexual–Transmission-Disease, Swelling-on-Neck, Different-Count, Total-Count, Erythrocyte-Rate, Creatinine, Loss-of-Appetite, Lymphodenopathy and the decision attribute Result (Positive, Negative, Suspect). In this study, decision attributes are omitted to analyze the proposed methodology, since there are possibilities to obtain the information system without decision attribute in the real life cases.

B) The K-Means Clustering Algorithm

In stage 2, the data set without decision attribute obtained from stage 1 is partitioned into K clusters, where each cluster comprises data-vectors with similar inherent characteristics. The overall outcome of this stage is the availability of K-number of data clusters, which forms the basis for subsequent discovery of symbolic rules that define the structure of the discovered clusters.

The K-Means Algorithm Process

[1] The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.
[2] For each data point, calculate the distance from the data point to each cluster.
[3] If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.
[4] Repeat the above step until a complete pass through all the data points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
[5] The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intra-cluster distances and cohesion.

C) Rule Extraction

In this stage, reduced data obtained from stage 3 is applied to the rule extraction algorithm to formulate the efficient rules (Table 6). The rule extraction algorithm uses the following Heuristic Approach:

(i) Merge identical rows that are rows with similar condition and decision attribute values.

(ii) Compute the core of every row.
(iii) Merge duplicate rows and compose a table with reduct value.

D) Worked example

A system of 8 data points consisting four condition attributes with no decision attribute is taken into consideration and it is presented in Table 1.

Table1: Data set

Object	Weight	Door	Size	Cylinder
1	LOW	2	COM	4
2	LOW	4	SUB	6
3	MEDIUM	4	COM	4
4	HIGH	2	COM	6
5	HIGH	4	COM	4
6	LOW	4	COM	4
7	HIGH	4	SUB	6
8	LOW	2	SUB	6

In Table 1, the following substitutions LOW=1, MEDIUM=2, HIGH=3, COM=1 and SUB=2 can be used. Applying K-Means Clustering algorithm with K=2. The clustered rows are {1, 3, 5, 6} and {2, 4, 7, 8}. Then the above table is reconstructed using the clustered rows as the decision value, presented in Table 2.

Table 2: Data set after K-means Clustering

Object	Weight	Door	Size	Cylinder	Mileage
1	1	2	1	4	1
2	1	4	2	6	2
3	2	4	1	4	1
4	3	2	1	6	2
5	3	4	1	4	1
6	1	4	1	4	1
7	3	4	2	6	2
8	1	2	2	6	2

Applying the Quickreduct algorithm in Table 2, the final reduct attributes {WEIGHT, DOOR, SIZE} is obtained. Hence, Table 2 can be reduced into Table 3 using the attribute reduct {WEIGHT, DOOR, SIZE}.

Table 3: Attribute Reduction

Object	Weight	Door	Size	Mileage
1	1	2	1	1
2	1	4	2	2
3	2	4	1	1
4	3	2	1	2
5	3	4	1	1

6	1	4	1	1
7	3	4	2	2
8	1	2	2	2

Rule Extraction:
Merge identical objects of Table 3. In this step, take the condition attributes of {WEIGHT, DOOR, SIZE} as presented in Table 3. If any identical pair occurs, merge it, otherwise compute the core of every object in Table 3 and present it as in Table 4.

Table 4: Core

Object	Weight	Door	Size	Mileage
1	1	*	1	1
2	1	*	2	2
3	*	4	1	1
4	3	*	*	2
5	*	4	1	1
6	1	*	1	1
7	3	*	*	2
8	1	*	2	2

In the next step, merge duplicate objects with same decision value and compose a table with the reduct value. That is, the merged rows are {1, 6}, {2, 8}, {3, 5} and {4, 7} as presented in Table 5.

Table 5: Merged Rows

Object	Weight	Door	Size	Mileage
1	1	*	1	1
2	1	*	2	2
3	*	4	1	1
4	3	*	*	2

Table 5 shows the new set of objects which contains the rules of Table 2. Decision rules are often presented as implications and are often called “if….then…” rules. We can express the rules as follows:

(i) If SIZE = 1 THEN MILEAGE = 1
(ii) If SIZE = 2 THEN MILEAGE = 2
(iii) If DOOR = 4 and SIZE = 1 THEN MILEAGE = 1
(iv) If WEIGHT = 3 THEN MILEAGE = 2

EXPERIMENTAL ANALYSIS

The K-Means Clustering, Quickreduct, VPRS and Rule extraction algorithm have been implemented using MATLAB for databases available in the UCI data repository and the HIV data directly collected from the 500 HIV patients. The Comparative Analysis of Quickreduct and VPRS is tabulated in Table 6 as given below. It is observed that less number of rules are generated for the reduct set obtained by using VPRS than the reduct set generated by using Quickreduct.

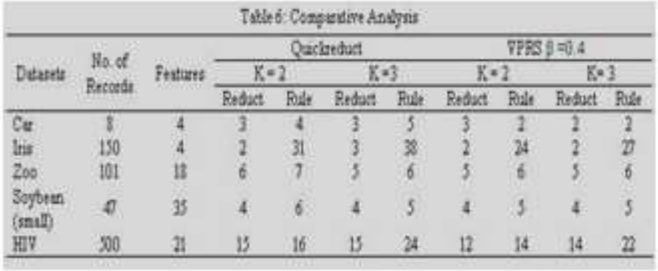

Table 6: Comparative Analysis

Datasets	No. of Records	Features	Quickreduct				VPRS β =0.4			
			K = 2		K =3		K = 2		K= 3	
			Reduct	Rule	Reduct	Rule	Reduct	Rule	Reduct	Rule
Car	8	4	3	4	3	5	3	2	2	2
Iris	150	4	2	31	3	38	2	24	2	27
Zoo	101	18	6	7	5	6	5	6	5	6
Soybean (small)	47	35	4	6	4	5	4	5	4	5
HIV	500	21	15	16	15	24	12	14	14	22

CONCLUSION

In the rule extraction process, almost all the researchers have framed the rules after applying any one of the reduct algorithms based on rough set theory approach or statistical approach. In this paper, the K-means algorithm has been used to cluster the data set into K-clusters. Then applying the Quick and VPRS reduct algorithms to get the best reduct set of attributes, it was found that the VPRS produces the best reduct for the large data set. The VPRS generates the reduct set which consists of 12 and 14 attributes for K = 2 and 3 respectively, whereas Quickreduct generates a reduct set with 15 attributes for K = 2 and K = 3 in the case of HIV data set. It was observed that less number of rules were produced when the VPRS reduct applied for K = 3 compared to Quickreduct. The unsupervised technique was applied for clustering in this work. The proposed work can be improved by introducing the Neural Network in order to train the system and this is the direction for further research work.

REFERENCES

[1] H. Almuallim and T.G.Dietterich. Learning with many irrelevant features. *9th National Conference on Artificial Intelligence, MIT Press*, 547-552, 1991.

[2] J.J. Alpigini, J.F. Peters, J. Skowronek and N.Zhong (Eds.). Rough Sets and Current Trends in Computing. *Third International Conference, RSCTC, Malvern, PA, USA*, 2002.

[3] M. J. Beynon. An investigation of β-reduct selection within the variable precision rough set model. *Proceedings of Second International Conference on Rough Sets and Current Trends in Computing*, 114-122, 2000.

[4] M.J. Beynon. Reducts within Variable Precision Rough Sets Model: A Further Investigation. *European Journal of Operational Research*, 134(3):592-605, 2001.

[5] C. L. Blake and C. J. Merz. UCI Repository of machine learning databases. *Irvine, University of California*, 1998, http://www.ics.uci.edu/~mlearn/.

[6] P. Dayan. Unsupervised learning. *The MIT Encyclopedia of the Cognitive Science*, 1999.

[7] J.G.Dy and C.E. Brodley Feature selection for unsurprised learning. Journal of Machine Learning Research,5,2004

[8] M.Goebel and L. Gruenwal. A Datamining and Knowledge Discovery Software Tools. SIGKDD Explorations, 11(1),1999.

[9] X. Hu, T.Y. Lin and J. Jianchao. A New Rough Sets Model Based on Database Systems. *Fundamenta Informaticae*, 1-18, 2004.

[10] R.Jensen and Q.Shen. A Rough Set – Aided system for Sorting WWW Book-marks. In N.Zhong et al.(Eds.), *Web Intelligence: Research and Development*, 95-105, 2001.

[11] R. Jensen. *Combining Rough and Fuzzy Sets for Feature Selection.* Ph.D Thesis, School of Informatics, University of Edinburgh, 2005.

[12] R. Jensen and Q. Shen. Fuzzy-rough attribute reduction with application to web categorization. *Fuzzy Sets and Systems*, 141(3): 469-485, 2004.

[13] R. Jensen and Q. Shen. Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approaches. *IEEE Transactions on Knowledge and Data Engineering*, 16(12), 2004.

[14] G.H. John, R. Kohavi and K. Pfleger. Irrelevant Features and the Subset Selection Problem. *Proceedings of 11th International Conference on Machine Learning*, 121-129, 1994.

[15] K. Kira and L.A. Rendell. The Feature selection Problem: Traditional Methods and a New Algorithm. *Proceedings of AAAI, MIT Press*, 129-134, 1992.

[16] R.Kohavi. Useful feature subsets and Rough set reducts. *Proceedings of the 3rd International Workshop on*

Rough Set and Soft Computing, 310-317, 1994.

[17] M. Kryszkiewicz. Maintenance of reducts in the variable precision rough sets model. *ICS Research Report, 31/94, Warsaw University of Technology*, 1994.

[18] A. Kusiak. Rough Set Theory: A Datamining Tool for Semiconductor Manufacturing. *IEEE Transactions on Electronics Packaging Manufacturing*, 24(1), 2001.

[19] T.Y. Lin, N. Cercone (Eds.). *Rough sets and Datamining: Analysis of Imprecise Data*. Kluwer Academic Publishers, 1997.

[20] H. Liu and H. Motoda (Eds.). *Feature Extraction Construction and Selection: A Datamining Perspective*, Kluwer International Series in Engineering and Computer Science Kluwer Academic Publishers, 1998.

[21] H.Liu and R.Setiono. A probabilistic approach to feature selection – a filter solution. *Proceedings of the 9th International conference on Industrial and Engineering Applications of AI and ES*, 284-292, 1996.

[22] H.Liu and R.Setiono. Feature selection and classification – a probabilistic wrapper approach. *Proceedings of the 9th International Conference on Industrial and Engineering Applications of AI and ES*, 419-424, 1996.

[23] T.Maciag and D.H.Hepting. Analysis of User Classifiers for Personalization of Environmental Decision Support System Interfaces. *Artificial Neural Networks in Engineering, St.Louis, Missouri*, 2005.

[24] T.Maciag, D.H.Hepting and D.Slezak. Personalizing User Interfaces for Environmental Decision Support Systems. *International Conference on Web Intelligence-Rough Set Workshop, Compiegne, France*, 2005.

[25] M. Modrzejewski. Feature Selection Using Rough M. Modrzejewski. Feature Selection Using Rough Sets Theory.