Available online at www.isas.org.in
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(2) 2009 181-188

Post Processing of Clusters for Pattern Discovery: Rough Set Approach

Alka Arora[1*], Shuchita Upadhyaya[2] and Rajni Jain[3]
[1]*Indian Agricultural Statistics Research Institute, New Delhi*
[2]*Kurukshetra University, Kurukshetra*
[3]*National Center for Agricultural Economics and Policy Research, New Delhi*

(Received: February 2008, Revised: May 2009, Accepted: May 2009)

SUMMARY

Most of clustering algorithms generate clustering results in the form of number of clusters and member objects in those clusters. This further requires analysis by experts in order to understand the patterns of obtained clusters. Post processing of cluster is then required in order to extract meaningful cluster pattern. In this paper a rough set based approach for pattern discovery from individual clusters is proposed. In the proposed approach, Maximum Possible Combination Reduct (MPCR) derived from rough set theory is used for generating concise cluster pattern. MPCR is defined as the set of variables which distinguishes the objects in a homogenous cluster. Therefore these variables are not considered for pattern formulation. Remaining variables are ranked for their contribution in the cluster. Cluster pattern is formed by conjunction of variables in the increasing order of their contribution in the cluster such that pattern distinctively describes the cluster with minimum error. Applicability of approach is demonstrated using soybean disease and zoo datasets from machine learning repository.

Keywords: Clustering, Data mining, Rough set theory, Reduct, Indiscernibility, MPCR, Cluster description, Pattern.

1. INTRODUCTION

Data Mining is a non trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data (Han and Kamber 2006). Clustering is an important component of data mining. The underlying assumption of clustering in data mining is to find out the hidden patterns in the data, which can be revealed by grouping the objects into clusters. According to Mirkin (2005), clustering process involves different stages, which include data pre-processing and standardization, finding clusters in data and description of clusters. Many clustering algorithms are available in literature, one can refer to Jain *et al*. 1999; Han and Kamber 2006; Mirkin 2005; for comprehensive surveys on clustering algorithms. *K*-Means and Expectation Maximization (EM) algorithms are the widely known partitional algorithms, which divide the data into k non overlapping clusters. These clustering algorithms just generate general description of the clusters like which objects are member of each cluster and lacks in generating cluster description in terms of relevant variables those define the cluster. According to Ganter and Wille (1997), cluster description is able to approximately describe the cluster in the form that "this cluster consists just of all the objects having the pattern P, where pattern is formulated using the variable and values of the given many valued context". From an intelligent data analysis perspective deriving knowledge in the form of pattern from obtained clusters is as important as grouping the objects into clusters.

[*] *Corresponding author* : Alka Arora
E-mail addresses : alkak@iasri.res.in, alka27@yahoo.com

Rough Set Theory (RST) proposed by Pawlak (1991), has been successfully applied in classification techniques for pattern/knowledge discovery (Komorowski 1999). RST has also relevance in clustering as RST divides the data into equivalence/ indiscernible classes; each indiscernible class can be considered as natural cluster. Moreover, RST performs automatic concept approximation by producing minimal subset of variables (Reduct) which can distinguish indiscernible classes in the dataset. In general, classification problems using rough sets involve computation of decision relative reduct. Clustering, an unsupervised method of data mining requires reduct computation purely on the basis of indiscernibility as there is no decision variable. Such reduct are referred as unsupervised reduct in this paper.

The proposed approach of cluster description is applied as post processing step on obtained clusters. As our aim is to generate characteristics of individual clusters, hence partition based algorithm is used to obtain non overlapping clusters. Applicability of proposed approach is studied on soybean disease and zoo datasets of agriculture domain from UCI repository (UCI). Objective of applying the proposed approach on soybean dataset is to study the relevant variables which contribute towards the occurrence of a particular disease and on zoo dataset is to characterize the animal clusters.

The paper is organized in six sections. Section 2 provides overview of rough set concepts. Section 3 gives background and related work in the area of cluster description. Section 4, provides the details of proposed approach. Section 5 details the application of proposed approach on soybean disease and zoo datasets followed by conclusions in Section 6.

2. ROUGH SET THEORY : A BRIEF OVERVIEW

RST is a mathematical approach, proposed by Pawlak (1991), further refined by Komorowski and Polkowski (1999), Yao *et al.* (1997), to cope with data analysis in the presence of imprecision, vagueness and uncertainty. In RST, dataset is represented in the form of information table; each case represents an object and columns represent variables. More formally it is an information system $X = (U, A)$ where U is non-empty, finite set of objects called the universe and A is non-empty, finite set of variables on U. With every variable $a \in A$, a set V_a is associated such that $a : U \rightarrow V_a$. The set V_a is called the domain or value set of variable a. Small table from soybean disease dataset is used for illustration (Table 1). The dataset has ten objects characterized by eight nominal variables.

2.1 Indiscernibility Relation

Indiscernibility relation is core concept of RST. Indiscernibility relation $IND(B)$, for any subset $B \subseteq A$ is defined by

$$IND(B) = \{(x, y) \mid a(x) = a(y),\ \forall\ a \in B\ ;\ x, y \in U\}$$

Two objects are considered to be indiscernible or similar by the variables in B, if and only if they have

Table 1. Small soybean dataset

id	date	precip	damage	severity	canker_lesion	fruiting_bodies	decay
X1	july	lt-norm	scattered	pot-severe	tan	absent	absent
X2	october	norm	scattered	pot-severe	tan	absent	absent
X3	september	lt-norm	whole-field	pot-severe	tan	absent	absent
X4	august	norm	whole-field	pot-severe	tan	present	absent
X5	august	lt-norm	upper-area	pot-severe	tan	absent	absent
X6	september	gt-norm	whole-field	pot-severe	dk-brown-blk	absent	absent
X7	july	gt-norm	scattered	pot-severe	dk-brown-blk	absent	firm-and-dry
X8	august	gt-norm	low-areas	pot-severe	dk-brown-blk	absent	firm-and-dry
X9	september	gt-norm	upper-area	minor	dk-brown-blk	absent	firm-and-dry
X10	october	gt-norm	whole-field	minor	dk-brown-blk	absent	firm-and-dry

the same value for every variable in B. Objects in the information system which have the same value form an equivalence relation. Equivalence relation partition set of objects (U) into set of equivalence classes. $IND(B)$ is an equivalence relation which partitions U into set of partitions denoted by $U / IND(B)$.

For example from Table 1, when B = {damage} then objects X1, X2 and X7 are indiscernible and therefore form one equivalence class; X3, X4, X6 and X10 are indiscernible and X5 is indiscernible with X9. Formally:

$$U/IND\{damage\} = \{\{X1,X2,X7\}, \{X3,X4,X6,X10\}, \{X5,X9\},\{X8\}\}$$

Similarly $U/IND\{canker\_lesion, decay\}$

$$= \{\{X1,X2,X3,X4,X5\},\{X6\}, \{X7,X8,X9,X10\}\}$$

2.2 Reduct

Concept approximation is achieved in RST through data reduction i.e. by retaining the minimum subset of variables that can differentiate all equivalence classes in the universe set. Such minimum subset is called reduct. More formally reduct R is a set of variables such that

$$R \subseteq A$$

$$IND_R(U) = IND_A(U)$$

$$IND_{R-a}(U) \neq IND_A(U) \qquad \forall\ a \in R$$

There are many methods as well as many software's available for computation of reduct, discussion on those is beyond the scope of this paper. We have considered Genetic Algorithm (GA) (Wroblewski 1995) for reduct computation, as it can produce many reducts of varying cardinality. This provides flexibility to the experimenter for selection of variables from the reduct population produced by GA. There are many approaches to consider variables from reducts generated by GA (Komorowski and Polkowski 1999). Maximum Possible Combined Reduct (MPCR) is defined as the union of variables present in the reduct sets obtained after applying GA (Jain 2004). Any variable that belongs to at least one of the reduct in the population of reducts from GA also belongs to MPCR. More formally MPCR is set of variables M, such that

$$M \subseteq A$$

$$M = \bigcup_{i=1}^{n} R_i$$ where R_i is the i^{th} reduct in the population of reducts from GA.

For Example, reduct computation on the Table 1 resulted in six reducts of cardinality three; R1 = {date, damage, canker_lesion}, R2 = {precip, damage, severity}, R3 = {date, precip, severity}, R4 = {date, precip, damage}, R5 = {date, precip, decay} and R6 = {precip, damage, decay}. MPCR set computation from these reducts is {date, precip, severity, damage, decay, canker_lesion}.

3. BACKGROUND AND RELATED WORK

Cluster description is useful in studying the object variable relationship which describes the underlying cluster. This can be applied in various areas for understanding the clusters viz. In disease diagnostic system, where there is a need to study the diseases characteristics; In Web Mining, finding pattern in the set of web users; Given a set of tourist places, finding out what features of places and tourist attract each other; In banks, customer data is available on many variables, discovery of age and salary as sufficient variables to grant loan to a customer; In characterization of animal and plant taxonomy clusters.

3.1 Review of Literature

In the literature, Mirkin (2005), Han and Kamber (2001), the problem of conceptual description of partition has received by far more attention than the problem of description of a single cluster. Decision tree is mainly used for conceptual description of partition as it provides easily understandable description. Primary goal of building a decision tree is prediction of the partition under consideration rather than its description. Limitation of this technique is that it is 'monothetic' and hence each split goes along with only one variable, and not directly applicable to cluster whose definition involve combination of variables. In clustering, the criterion is to get clusters as homogenous as possible with regard to all the variables however in decision tree; criterion is homogeneity with regard to a pre-specified decision variable.

As discussed by Mirkin (2005), the problem of producing description for a single cluster without any

relevance to other clusters has recently attracted considerable attention from the researchers. There are few references of cluster description approaches available in literature. Mirkin (1999) has proposed a method for cluster description applicable to only continuous variables. In Mirkin's approach variables are normalized first and then ordered according to their contribution weights which are proportional to the squared differences between their within group averages and grand means. A conjunctive description of cluster is then formed by consecutively adding variables according to the sorted order. Description is evaluated on precision error. Abidi *et al*. (1998, 2001) has proposed the rough set theory based method for rule creation for unsupervised data using dynamic reduct. Dynamic reduct is defined as the frequently occurring reduct set from the samples of original decision table. However, these approaches have their limitations. Mirkin's (1999) approach is applicable only to datasets having continuous variables. Abidi *et al*. (1998, 2001) in his approach has used the cluster information obtained after cluster finding and generated rules from entire data with respect to cluster/class attribute, instead of producing description for individual clusters. However, our approach is to generate user understandable cluster description for individual clusters by conjunction of significant variables which define the cluster.

3.2 Cluster Description Evaluation Criteria

As discussed by Mirkin (1999), accuracy of obtained pattern is measured in terms of Precision Error (*PE*). *PE* of pattern *P*, *PE* (*P*) is defined as

$$PE(P) = \frac{\left| false\ positive\ C(P) \right|}{\left| U - C \right|} \qquad (1)$$

where numerator, *false positive C(P)* is defined as the number of objects that lies outside cluster *C*, for which pattern *P* is true and denominator denotes the number of objects outside *C*.

4. PROPOSED APPROACH (REDUCT DRIVEN CLUSTER DESCRIPTION–RCD)

Proposed pattern discovery approach for individual clusters, called RCD is applicable as post processing step to clusters obtained using partition based clustering algorithm. RCD approach is divided into three stages.

4.1 Cluster Finding

First stage deals with obtaining clusters from dataset by applying clustering algorithm. We have used Weka implementation of EM algorithm for cluster finding (Weka). EM models the distribution of the objects probabilistically, so that an object belongs to a cluster with certain probability. The first step, calculation of the cluster probabilities, which are the expected class value, is "expectation"; the second step which deals with calculation of the distribution parameter is "maximization" of the likelihood of the distribution given the data (Mirkin 2005).

We have selected EM algorithm as it can handle both numeric and nominal variables. Weka implementation of EM algorithm has built in evaluation measure for computing the number of clusters present in the dataset. EM selects the number of clusters automatically by maximizing the logarithm of the likelihood of future data, estimated using cross-validation. Beginning with one cluster, it continues to add clusters until the estimated log-likelihood decreases (Weka).

4.2 Computation of Unsupervised Reduct

Clustering algorithm is intended to form clusters having most variable values common to their members (cohesion) and few values common to members of other clusters (distinctiveness) (Talavera 1999). Hence, variables which have similar value for majority of objects in the cluster are considered significant and rest are non significant for generating cluster pattern (Arora 2007).

Reduct accounts for discerning between the objects in a cluster, hence computation of unsupervised reduct in a cluster C provides the set of non significant variables for that cluster. Genetic Algorithm produces many reducts, hence computation of MPCR set (*RC*) in a cluster *C* provides the set of non-significant variables for that cluster. These non-significant variables (reduct) can be straight away removed from the cluster. The remaining variables (non reduct) form the set of significant variables (I) for that cluster.

4.3 Cluster Description

Cluster description approximately describes the cluster in the form of pattern. Pattern is formulated by

conjunction of significant attribute = value pairs from that cluster. There can be many possible patterns for a single cluster. Our aim is not to generate all possible patterns, but meaningful and concise pattern from the cluster. Therefore attributes in set are then ranked on Precision Error (*PE*) which is defined as

$$PE\ (a = v) = \frac{\left|\ false\ \ positive\ \ C(a=v)\ \right|}{\left|\ U-C\ \right|} \qquad (2)$$

where numerator defines the number of entities that lies outside cluster C, for which $a = v$ ($a \in A$, $v \in V_a$) is true and denominator defines the number of entities outside cluster C. An attribute value pair $a = v$ is said to be more contributing if it has less *PE*, means majority of objects satisfying this attribute value pair belongs to a single cluster.

Therefore problem of cluster description can be defined as forming a description P by combining the significant variables with less *PE* such that *PE* for P is minimum. Hence pattern P distinctively describes the cluster.

Procedure for RCD approach

1. Obtain clusters by applying partitional clustering algorithm.
2. Compute unsupervised reduct for individual clusters and then compute MPCR set (*RC*) for every cluster C.
3. Compute set of significant variables (I) for C, where $I = A - RC$.
4. Calculate *PE* for significant variables in set I for cluster C and arrange the set I in increasing order of *PE* score.
5. Combine variables from I with less *PE* to make the description such that *PE* for that description is minimum.

5. EXAMPLE OF APPLICATION

In this section, we illustrate the application of RCD approach on soybean disease dataset, followed by results of the same on Zoo dataset from UCI repository.

5.1 Soybean Dataset

In soybean disease set, Universal set (U) contains 47 objects and set of variables (A) consist of 35 multi-valued variables characterizing diaporthe-stem-canker, charcoal-rot, rhizoctonia-root-rot and phytophthora-rot diseases. All the variables are nominal in nature. Variables are broadly categorized into environmental descriptors, condition of leaves, condition of stem, condition of fruit pods and condition of root. Table 2 shows variable information of soybean dataset. It is observed that dataset is having unique value for some of the variables hence those variables are irrelevant and removed from the dataset. Reduced dataset then has 20 variables characterizing soybean

Table 2. Variable information of soybean dataset

v1	date: april=0, may=1, june=2, july=3, august=4, september=5, october=6
v2	plant-stand: normal=0, lt-normal=1
v3	precip: lt-norm=0, norm=1, gt-norm=2
v4	temp: lt-norm=0, norm=1, gt-norm=2
v5	hail: yes=0, no=1
v6	crop-hist: diff-lst-year=0, same-lst-yr=1, same-lst-two-yrs=2, same-lst-sev-yrs=3
v7	area-damaged: scattered=0, low-areas=1, upper-areas=2, whole-field=3
v8	severity: pot-severe=1, severe=2
v9	seed-tmt: none=0, fungicide=1
v10	germination: '90-100%'=0, '80-89%'=1, 'lt-80%'=2
v11	plant-growth: abnorm=1
v12	leaves: norm=0, abnorm=1
v13	leafspots-halo: absent=0
v14	leafspots-marg: dna=2
v15	leafspot-size: dna=2
v16	leaf-shread: absent=0
v17	leaf-malf: absent=0
v18	leaf-mild: absent=0
v19	stem: abnorm=1
v20	lodging: yes=0, no=1
v21	stem-cankers: absent=0, below-soil=1, above-soil=2, above-sec-nde=3
v22	canker-lesion: dna=0, brown=1, dk-brown-blk=2, tan=3
v23	fruiting-bodies: absent=0, present=1
v24	external decay: absent=0, firm-and-dry=1
v25	mycelium: absent=0, present=1
v26	int-discolor: none=0, black=2
v27	sclerotia: absent=0, present=1
v28	fruit-pods: norm=0, dna=3
v29	fruit spots: dna=4
v30	seed: norm=0
v31	mold-growth: absent=0
v32	seed-discolor: absent=0
v33	seed-size: norm=0
v34	shriveling: absent=0
v35	roots: norm=0, rotted=1

diseases. Dataset consist of instance number and class variables that are not considered while clustering.

EM clustering algorithm learnt four clusters from the dataset. Table 3 shows the dataset along with cluster information.

Table 3. Soybean dataset with clustering results

Sno	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v12	v20	v21	v22	v23	v24	v25	v26	v27	v28	v35	Cluster
0	4	0	2	1	1	1	0	1	0	2	1	0	3	1	1	1	0	0	0	0	0	cluster1
1	5	0	2	1	0	3	1	1	1	2	1	1	3	0	1	1	0	0	0	0	0	cluster1
2	3	0	2	1	0	2	0	2	1	1	1	0	3	0	1	1	0	0	0	0	0	cluster1
3	6	0	2	1	0	1	1	1	0	0	1	1	3	1	1	1	0	0	0	0	0	cluster1
4	4	0	2	1	0	3	0	2	0	2	1	0	3	1	1	1	0	0	0	0	0	cluster1
5	5	0	2	1	0	2	0	1	1	0	1	1	3	1	1	1	0	0	0	0	0	cluster1
6	3	0	2	1	0	2	1	1	0	1	1	1	3	0	1	1	0	0	0	0	0	cluster1
7	3	0	2	1	0	1	0	2	1	2	1	0	3	0	1	1	0	0	0	0	0	cluster1
8	6	0	2	1	0	3	0	1	1	1	1	0	3	1	1	1	0	0	0	0	0	cluster1
9	6	0	2	1	0	1	0	1	0	2	1	0	3	1	1	1	0	0	0	0	0	cluster1
10	6	0	0	2	1	0	2	1	0	0	1	1	0	3	0	0	0	2	1	0	0	cluster2
11	4	0	0	1	0	2	3	1	1	1	1	0	0	3	0	0	0	2	1	0	0	cluster2
12	5	0	0	2	0	3	2	1	0	2	1	0	0	3	0	0	0	2	1	0	0	cluster2
13	6	0	0	1	1	3	3	1	1	0	1	0	0	3	0	0	0	2	1	0	0	cluster2
14	3	0	0	2	1	0	2	1	0	1	1	0	0	3	0	0	0	2	1	0	0	cluster2
15	4	0	0	1	1	1	3	1	1	1	1	1	0	3	0	0	0	2	1	0	0	cluster2
16	3	0	0	1	0	1	2	1	0	0	1	0	0	3	0	0	0	2	1	0	0	cluster2
17	5	0	0	2	1	2	2	1	0	2	1	1	0	3	0	0	0	2	1	0	0	cluster2
18	6	0	0	2	0	1	3	1	1	0	1	0	0	3	0	0	0	2	1	0	0	cluster2
19	5	0	0	2	1	3	3	1	1	2	1	0	0	3	0	0	0	2	1	0	0	cluster2
20	0	1	2	0	0	1	1	1	1	1	0	0	1	1	0	1	1	0	0	3	0	cluster3
21	2	1	2	0	0	3	1	2	0	1	0	0	1	1	0	1	0	0	0	3	0	cluster3
22	2	1	2	0	0	2	1	1	0	2	0	0	1	1	0	1	1	0	0	3	0	cluster3
23	0	1	2	0	0	0	1	1	1	2	0	0	1	1	0	1	0	0	0	3	0	cluster3
24	0	1	2	0	0	2	1	1	1	1	0	0	1	1	0	1	0	0	0	3	0	cluster3
25	4	0	2	0	1	0	1	2	0	2	1	1	1	1	0	1	1	0	0	3	0	cluster3
26	2	1	2	0	0	3	1	2	0	2	0	0	1	1	0	1	1	0	0	3	0	cluster3
27	0	1	2	0	0	0	1	1	0	1	0	0	1	1	0	1	0	0	0	3	1	cluster3
28	3	0	2	0	1	3	1	2	0	1	0	1	1	1	0	1	1	0	0	3	0	cluster3
29	0	1	2	0	0	1	1	2	1	2	0	0	1	1	0	1	0	0	0	3	0	cluster3
30	2	1	2	1	1	3	1	2	1	2	1	0	2	2	0	1	0	0	0	3	1	cluster4
31	0	1	1	1	0	1	1	1	0	0	1	0	1	2	0	0	0	0	0	3	1	cluster4
32	3	1	2	0	0	1	1	2	1	0	1	0	2	2	0	0	0	0	0	3	1	cluster4
33	2	1	2	1	1	1	1	2	0	2	1	0	1	2	0	1	0	0	0	3	1	cluster4
34	1	1	2	0	0	3	1	1	1	2	1	0	2	2	0	0	0	0	0	3	1	cluster4
35	1	1	2	1	0	0	1	2	1	1	1	0	2	2	0	0	0	0	0	3	1	cluster4
36	0	1	2	1	0	3	1	1	0	0	1	0	1	2	0	0	0	0	0	3	1	cluster4
37	2	1	2	0	0	1	1	2	0	0	1	0	1	2	0	0	0	0	0	3	1	cluster4
38	3	1	2	0	0	2	1	2	1	1	1	0	2	2	0	0	0	0	0	3	1	cluster4
39	3	1	1	0	0	2	1	2	1	2	1	0	2	2	0	0	0	0	0	3	1	cluster4
40	0	1	2	1	1	1	1	1	0	0	1	0	1	2	0	1	0	0	0	3	1	cluster4
41	1	1	2	1	1	3	1	2	0	1	1	1	1	2	0	1	0	0	0	3	1	cluster4
42	1	1	2	0	0	0	1	2	1	0	1	0	2	2	0	0	0	0	0	3	1	cluster4
43	1	1	2	1	1	2	3	1	1	1	1	0	2	2	0	1	0	0	0	3	1	cluster4
44	2	1	1	0	0	3	1	2	0	2	1	0	1	2	0	0	0	0	0	3	1	cluster4
45	0	1	1	1	1	2	1	2	1	0	1	1	2	2	0	1	0	0	0	3	1	cluster4
46	0	1	2	1	0	3	1	1	0	2	1	0	1	2	0	0	0	0	0	3	1	cluster4

In order to study the disease characteristics, reduct analysis is carried out on individual four disease clusters. Table 4 shows the MPCR variables in different clusters.

Table 4. MPCR variables in different clusters

	MPCR variables
Cluster1	v1, v5, v6, v7, v8, v9, v10, v20, v22
Cluster2	v1, v4, v5, v6, v7, v9, v10, v20
Cluster3	v1, v5, v6, v8, v9, v10, v12, v20, v25, v35
Cluster4	v1, v3, v4, v5, v6, v8, v9, v10, v21, v24

Reduct analysis on different clusters shows that it has different MPCR variables, as variables are having different values in different clusters. Variables are not common across clusters and as such some variables are playing role in one cluster and not in other cluster.

Let us consider cluster4 for illustration (Table 3). Cluster4 has 17 entities of phytophthora-rot disease. To study the disease characteristic, reduct analysis is carried out on this cluster. Reduct computation on this cluster resulted in 22 reducts of varying cardinality. MPCR set is then computed from these reducts. Removal of MPCR variables (v1, v3, v4, v5, v6, v8, v9, v10, v21, v24) (Table 4) resulted in cluster having same value for all of its instances. These remaining variables (v7 = 1, v12 = 1, v20 = 1, v22 = 2, v23 = 0, v25 = 0, v26 = 0, v27 = 0, v28 = 3, v35 = 1) are playing major role in characterizing this specific cluster. PE is calculated for these remaining variables. In Cluster4, PE for variable v7 = 1 is 13/30 (Equ. 2), as 3 entities from Cluster1 and 10 entities from Cluster3 are satisfying this condition (Table 3). Similarly PE for other variables in this cluster are v12 = 1(21/30), v20 = 1(21/30), v22 = 2(0), v23 = 0(20), v25 = 0(25), v26 = 0(20), v27 = 0(20), v28 = 3(10) and v35 = 1(1). PE for variable v22 = 2 is zero, hence variable v22 = 2 describes this cluster with no error.

Table 5. Patterns obtained for soybean disease clusters

Cluster	Pattern	PE
Cluster 1 (diaporthe-stem-canker)	stem-cankers = above-sec-nde or fruiting-bodies = present	0
Cluster 2 charcoal-rot	precip = lt-norm or stem-cankers = absent or canker-lesion = tan or int-discolor = black or sclerotia = present	0
Cluster 3 (rhizoctonia-root-rot)	canker-lesion = brown ^ temp = lt-norm	0
Cluster 4 (phytophthora-rot)	canker-lesion = dk-brown-blk	0

Let us consider another example of Cluster3 (Table 3) which has ten entities corresponding to disease rhizoctonia-root-rot. After the removal of MPCR variables (Table 4) (v1, v5, v6, v8, v9, v10, v12, v20, v25, v35) from this cluster, remaining variables (v3, v4, v7, v21, v22, v23, v24, v26, v27 and v28) are having same value for all of its instances. PE for these remaining variables are v3 = 2(23/37), v4 = 0(7/37), v7 = 1(19/37), v21 = 1(8/37), v22 = 1(6/37), v23 = 0(27/37), v24 = 1(16/37), v26 = 0(27/37), v27 = 0(27/37) and v28 = 3(17/37). There is no variable with PE zero, therefore as per proposed approach combine together the variables with less PE, v22 with PE 6/37and v4 with PE 7/37. Description P: v22 = 1 and v4 = 0 describes this cluster with zero error. Similarly for Cluster2 variables v3 = 0, v21 = 0, v22 = 3, v26 = 2 and v27 = 1 have zero PE, hence any of these variables can describe the cluster completely. Cluster1 have variables v21 = 3 and v23 = 1 with zero PE, hence either of these variables can describe the cluster without error. Results of cluster description on soybean disease clusters are summarizes in Table 5 (combining together name of the variables from Table 2):

5.2 Zoo Dataset

Zoo dataset consist of 101 instances of animals with 17 variables and 7 output classes (UCI). There are 15 boolean attributes, with value one and zero corresponding to the presence and absence of hair, feathers, eggs, milk, backbone, fins, tail, airborne, aquatic, predator, toothed, breathes, venomous, domestic and catsize. The attribute number of legs {0, 2, 4, 5, 6, 8}correspond to character variable. Variables animal name and class are not considered for clustering.

EM clustering algorithm learnt four clusters from the data instead of seven classes that is known in the dataset. Table 6 shows EM clustering results on Zoo dataset. Previous studies on clustering for zoo dataset and cluster validity indices also indicated better partitioning at two, four and seven clusters (Mitra *et al.* (2002)).

Table 6. EM clustering results on zoo dataset

Cluster Name	Cluster 0	Cluster 1	Cluster 2	Cluster 3
No. of objects	21	40	20	20

Unsupervised reduct is computed for individual clusters and then MPCR is computed from them. Table 7 shows MPCR variables in different clusters. Table 8 shows the results of cluster description for animal clusters.

Table 7. MPCR variables in individual clusters

Cluster	Reduct
Cluster 0	hair, airborne, predator, toothed, venomous, legs, domestic, backbone, breathes
Cluster 1	eggs, airborne, aquatic, predator, toothed, legs, tail, domestic, catsize
Cluster 2	airborne, aquatic, predator, domestic, catsize
Cluster 3	eggs, milk, aquatic, predator, breathes, venomous, legs, domestic, catsize

Table 8. Cluster description for animal clusters

Cluster	Number of elements in Cluster	Pattern	PE
Cluster 0	20	tail = 0 ^ milk = 0	0
Cluster 1	40	milk = 1 ^ hair = 1	0
Cluster 2	20	feathers = 1	0
Cluster 3	20	fins = 1	0.024

6. CONCLUSION

Clustering provides unsupervised grouping of objects in the form of clusters which needs to be analyzed and understood. In this paper, we presented reduct driven approach for selection of significant variables from individual clusters. Ranking of significant variables on precision error resulted in formulation of meaningful and concise cluster pattern. With the application of proposed approach on soybean and zoo datasets, it is observed that obtained patterns distinctively described the clusters with no or minimum errors. In future, RCD approach will be experimented with other datasets from different domains to study the effectiveness of this approach in generating cluster pattern.

ACKNOWLEDGEMENT

Authors are grateful to the anonymous referee for giving valuable suggestions that helped in significantly improving the quality of the paper.

REFERENCES

Arora, A., Upadhyaya, S. and Jain, R. (2007). Rough set approach for generating cluster description. *Proc. of the Information Systems, Technology and Management (ICISTM-2007)*, IMT Ghaziabad, ISBN: 81-8424-182-8, Allied Publishers Pvt. Ltd, 304-310.

Abidi, S.S.R., Hoe, K.M. and Goh, A. (2001). Analyzing data clusters: A rough set approach to extract cluster defining symbolic rules. *Proc. Fisher, Hand, Hoffman, Adams (eds.) Lecture Notes in Computer Science: Advances in Intelligent Data Analysis*, 4th Intl. Symposium, IDA-01. Springer Verlag, Berlin.

Abidi, S.S.R. and Goh, A. (1998). Applying knowledge discovery to predict infectious disease epidemics. *Proc. H. Lee and H. Motoda (eds.) Lecture notes in Artificial Intelligence 1531-PRICAI'98: Topics in Artificial Intelligence*, A Springer Verlag, Berlin.

Ganter, B. and Wille, R. (1997). *Formal Concept Analysis: Mathematical Foundations*. Springer- Verlag, New York Inc., Secaucus, New York.

Han, J. and Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999). Data clustering: A review. *ACM Computing Surveys*, **31(3)**, 264-323.

Jain, R. (2004). Rough Set based Decision Tree Induction for Data Mining. *Ph.D. Thesis*, JNU, New Delhi.

Komorowski, J., Pawlak, Z. and Polkowski, S. (1999). Rough sets: A tutorial. *In: S. K. Pal, A. Skowron (ed.). Rough Fuzzy Hybridization: A New Trend in Decision-Making*, Springer-Verlag, Berlin, 3–98.

Mirkin, B. (1999). Concept learning and feature selection based on square-error clustering. *Machine Learning*, **35**, 25-40.

Mirkin, B. (2005). *Clustering for Data Mining: Data Recovery Approach*. Chapman and Hall.

Mitra S., Pal, S.K., Mitra, P. (2002). Data Mining in Soft Computing Framework : A Survey, *IEEE Transactions on Neural Networks*, **13(1)**, 3-14.

Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers.

RSES: Rough Set Exploring System, available at: http://logic.mimuw.edu.pl/~rses.

Talavera, L. (1999). Feature selection as retrospective pruning in hierarchical clustering. *Proc. Third International Symposium on Intelligent Data Analysis*, IDA99 Amsterdam, Springer Verlag, The Netherlands.

UCI: Repository of Databases for Machine Learning and Data Mining, Urvine, UCI.

WEKA: A machine learning software available at: http://www.cs.waikato.ac.nz/~ml.