http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 8 (2): 173-180, 2009
ISSN 1812-5638

Integrated Approach of Reduct and Clustering for Mining Patterns from Clusters

[1]A. Arora, [2]S. Upadhyaya and [3]R. Jain
[1]Indian Agricultural Statistics Research Institute, Library Avenue,
Pusa, New Delhi-110012, India
[2]Department of Computer Science and Application,
Kurukshetra University, Kurukshetra, India
[3]National Center for Agricultural Economics and Policy Research,
Library Avenue, Pusa, New Delhi-110012, India

Abstract: In this study, a method is presented for selection and ranking of significant attributes for individual clusters which lead to formulation of concise and user understandable patterns. Cluster is set of similar data objects and similarity is measured on attribute values. Attributes which have same value for majority of objects in a cluster are considered significant and rest non significant for that cluster. Reduct from rough set theory is defined as the set of attributes which distinguishes the objects in a homogenous cluster, therefore these can be clear cut removed from the same. Non reduct attributes are ranked for their contribution in the cluster. Pattern is then formed by conjunction of most contributing attributes of that cluster.

Key words: Clustering, data mining, cluster description, pattern, rough set theory, reduct, indiscernibility

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 important component of data mining for grouping objects, such that objects within a single cluster have similar characteristics, hence objects in different clusters are dissimilar. For example, consider the disease dataset, which can be used to cluster different disease, such that objects of similar disease are in single cluster. The clusters can then be used to characterize the different diseases such that actions are directed towards specific clusters. There are significant clustering algorithms available in literature, broadly categorized in hierarchical and partitional categories. One can refer to Jain *et al.* (1999), Han and Kamber (2006) and Mirkin (2005), for comprehensive surveys on clustering algorithms. Majority of clustering algorithms generate general description of clusters like number of clusters, information about numerical similarity between objects and member objects of different clusters and lacks in generation of user understandable cluster description. According to Ganter and Wille (1997), conceptual cluster description is able to approximately describe the cluster in the form of pattern, where pattern is formulated by conjunction of significant attribute = value pair (descriptor) from that cluster and hence easily understood by the user.

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. Hence partitional clustering algorithm makes the obvious choice for our experiments as these algorithms divide the data into k non overlapping clusters. Post processing of clusters is required to extract meaningful mined patterns for the user. Meamingfulness of pattern is characterize on the measures; (1) pattern is supported by majority of objects from the cluster, (2) number of descriptors in pattern is small/less, (3) majority of objects satisfying the pattern belongs to a single cluster.

Rough Set Theory (RST) proposed by Pawlak (1995), has been successfully applied in classification techniques for pattern/knowledge discovery. RST has an appeal to be applied in clustering, as RST divides the data into indiscernible/equivalence classes; each indiscernible class can be considered as natural cluster. Moreover, RST performs automatic concept approximation by producing minimal subset of attributes (Reduct) which can distinguish indiscernible classes in the dataset. In general, classification problems using rough sets involve computation of decision/class relative reduct. Clustering, an unsupervised method of data mining requires reduct

Corresponding Author: Alka Arora, Indian Agricultural Statistics Research Institute, Library Avenue, Pusa, New Delhi-110012, India

computation purely on the basis of indiscernibility as there is no class/decision attribute. Such reducts are referred as unsupervised reduct in this study. Aim of the present study is to generate pattern of individual clusters; hence unsupervised reduct is computed for individual clusters in comparison to reduct computation for the dataset.

The proposed multi stage approach of cluster description involves (1) cluster formation using partional clustering algorithm (2) computation of unsupervised reduct (non significant attributes) for individual clusters (3) ranking of significant attributes (4) formulation of concise user understandable pattern.

Pattern discovery is useful in studying the object attribute relationship which describes the underlying cluster. This can be applied in various areas for understanding the pattern 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 attributes, discovery of age and salary as sufficient attributes to grant loan to a customer; In agriculture, for characterization of animal and plant taxonomy clusters.

The field of producing cluster description for individual clusters is relatively new. There are few references of cluster description approaches available in literature. Mirkin (1999) has proposed a method for cluster description applicable to only continuous attributes. In Mirkin's approach attributes are normalized first and then ordered according to their contribution weights which are proportional to the squared differences between their with-in group averages and grand means. A conjunctive description of cluster is then formed by consecutively adding attributes 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. However, they have 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.

Proposed approach presents an alternative method for formulation of short and user understandable patterns for individual clusters. Pattern for individual clusters are formulated by conjunction of significant attributes which define the cluster.

ROUGH SET THEORY

In RST (Pawlak *et al.*, 1995; Polkowski, 2002; Skowron *et al.*, 2002) data is represented as an information system $X = (U, A)$. In this U is non-empty finite set of objects called the universe and A is a non-empty, finite set of attributes on U. With every attribute $a \in A$, 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 attribute a. Every object x, in an information system x, is characterized by its information vector, $InfX(x) = \{(a, a(x): a \in A\}$.

Indiscernibility relation: Indiscernibility relation is core concept of RST. Indiscernibility relation IND (B), for any subset $B \subseteq A$ is defined as $x\ IND\ (B)\ y \Leftrightarrow \forall_{a \in B}(a(x) = a(y))$. Two objects are considered to be indiscernible/similar by the attributes in B, if and only if they have the same value for every attribute in B. Objects in the information system that share same knowledge form an equivalence relation. IND (B) is an equivalence relation that partitions U into set of equivalence classes. Set of such partitions are denoted by U/IND (B).

Reduct: An equivalence relation induces a partitioning of the universe. In general reduct are the minimum subset of attributes that can differentiate all equivalence classes in the universe. In the present study, reduct is considered as the set of attributes which distinguishes objects in an equivalence class/cluster as compared to all equivalence classes in the universe set.

There are many methods as well as many software's available for computation of reduct, discussion on those is beyond the scope of this study. We have considered Genetic Algorithm (GA) (Wróblewski, 1995) for reduct computation as it can produce sufficiently many reduct sets of varying length in considerable time. Rosetta software (Öhrn *et al.*, 1998) is used for computation of unsupervised reduct using GA.

Significance of attributes: Reduct set may contain more than one attribute. There are many approaches to consider relevant attributes from the reduct sets produced by GA. One approach is to consider core attributes i.e., common attributes shared by all the reduct sets. Another approach is to consider union of attributes present in the reduct sets i.e., Maximum Possible Combination Reduct (MPCR) (Minz and Jain, 2005). Another approach is related to dynamic reducts i.e., attribute sets appearing sufficiently often as reduct of samples of the original decision table. The attributes belonging to the majority of dynamic reducts are defined as relevant. It is also possible that GA may result in single reduct set with all the attributes. In this situation, approximate solution for reduct computation is considered in present study. Approximate solution for reduct computation covers enough sets from the function of reduct computation. On the lines of relevant attributes from dynamic reduct, in the present study reduct attributes belonging to majority of

approximate reduct sets produced by GA are considered relevant. These reduct attributes are defined as dynamic reduct attributes in this study. The value threshold for enough and majority needs to be tuned for the given data.

PROPOSED APPROACH OF CLUSTER DESCRIPTION

Here, proposed Reduct driven approach for Cluster Description (RCD) is described. RCD approach is divided into four stages.

Cluster finding: First stage deals with obtaining clusters from dataset by applying partional clustering algorithm. Weka implementation (Holmes *et al*., 1994) of EM algorithm is used for cluster finding. EM models the distribution of the objects probabilistically, so that an object belongs to a cluster with certain probability (Mirkin, 2005). The first step, calculation of the cluster probabilities, which are the expected class value, is expectation; the second step is calculation of the distribution parameter is maximization of the likelihood of the distribution given the data (Mirkin, 2005).

EM algorithm can handle different types of attributes. 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.

Computation of unsupervised reduct: Clustering algorithm is intended to form clusters having most attribute values common to their members (cohesion) and few values common to members of other clusters (distinctiveness) (Han and Kamber, 2006). Given a cluster C_i and a certain value v for attribute a, majority of cluster members will exhibit this value (Talavera, 1999). Intuitively, attributes which have similar value for majority of objects in the cluster are considered significant and rest are non significant for that cluster.

In the second stage, sets of significant and non significant attributes are computed for individual clusters. Unsupervised reduct is computed for individual clusters. Reduct accounts for discerning between the objects in a cluster, hence computation of reduct set (RC_i) in a cluster C_i provides the set of non significant attributes. These non significant attributes (reduct) can be straight away removed from the cluster. Remaining attributes (non reduct) form the set of significant descriptors (I) for that cluster. Formally, in the information system, variable value pair of the form $\{(a = v); a \in A, v \in V_a\}$ is defined as descriptor d (Ganter and Wille, 1997).

Ranking of significant attributes: Pattern formed with the conjunction of all significant descriptors can be quite complex and for maximum comprehensibility, shorter cluster description is preferred (Mirkin, 2005). Hence for generating short cluster description, in the third stage, descriptors ($dj \in I$; $j = 1,..., m$) are evaluated for their degree of significance in the cluster. Let $[a = v]_U$ and $[a = v]_{Ci}$ denote the set of objects satisfying $(a = v)$ in whole dataset and in cluster C_i, respectively. Let sup $port_U$ $(a = v) = card([a = v]_U)$ = number of objects in universe set and sup $port_{Ci}$ $(a = v) = card([a = v]_{Ci})$ number of objects in cluster C_i. A descriptor is said to be highly significant, if all the objects satisfying that condition belongs to a single cluster. It is quite possible that some objects that satisfy the descriptor also belongs to other clusters. Hence ranking of descriptors are proposed for evaluating their significance in the cluster. Descriptors are ranked on Precision Error (PE(d)), which is defined as:

$$PE(d) = \frac{\left|\text{sup port}_U(a = v) - \text{sup port}_{C_i}(a = v)\right|}{\left|\text{card } U - \text{card } C_i\right|} \quad (1)$$

If a descriptor has zero PE, which means all the objects satisfying that descriptor belongs to a single cluster. This is quite possible that a descriptor may have zero PE, but is not supported by majority of objects of that cluster. Hence another measure Coverage Ratio in Cluster (CRC(d)) is considered, which is defined as:

$$CRC(d) = \frac{\text{card}_{C_i}(d)}{\text{card } C_i} \quad (2)$$

Single measure of PE(d) or CRC(d) is not sufficient in determining degree of significance for descriptorof. Hence new measure Precision Coverage Coefficient (PCC(d)) is introduced, which is defined as:

$$PCC(d) = SQRT((1 - PE) * CRC) \quad (3)$$

Descriptors ($dj \in I$; $j = 1,..., m$) of C_i are evaluated on PCC score, which gives a real number in the interval [0, 1]; that expresses the significance of descriptor in a cluster. Descriptors in set I are then arranged in chronological order of PCC score.

Pattern generation: Pattern P_i of cluster C_i can be defined as:

$$P_i = \mathop{A}_{j=1}^{m} d_j ; d_j \in I$$

which is formed by concatenating significant descriptors from cluster C_i. A single descriptor is sufficient to

describe the cluster completely if it has the PCC score 1; otherwise pattern is formed by conjunction of descriptors, starting from highest to lowest value of PCC score. 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. Hence, descriptors satisfying PCC threshold (λ) are selected for pattern generation. Descriptor selection threshold (λ) can be application dependent and specified by the user.

It is quite possible that some objects that do not belongs to C_i, also satisfies P_i. Therefore, as defined by Mirkin (1999), pattern is evaluated on Precision Error (PE). Precision Error of pattern P_i, PE (Pi) is defined as:

$$PE(P_i) = \frac{\left|support_U(P_i) - support_{C_i}(P_i)\right|}{\left|card\ U - card\ C_i\right|} \quad (4)$$

This is also feasible that a pattern may have zero PE, but will not be able to cover maximum objects in a cluster. Therefore another measures of Coverage Ratio in Cluster CRC(Pi) and Pattern Length (L(Pi)) are considered for pattern evaluation, which are defined as:

$$CRC(P_i) = \frac{card_{C_i}(P_i)}{card\ C_i} \quad (5)$$

$$L(P_i) = \text{Number of descriptors occuring in } P_i \quad (6)$$

A pattern is said to be meaningful, if it has minimal PE, maximum coverage and less number of descriptors. At times it is not possible to obtain cluster description without any error and with full coverage. Hence, user specified threshold α of PE and β of CRC is input to algorithm, to generate pattern satisfying those parameters. An algorithm CLUSTDESC has been proposed for generating concise pattern. Steps for RCD are summarized here:

Step 1: Data clustering: Apply EM clustering algorithm on dataset to obtain clusters

Step 2: Reduct computation: Compute unsupervised reduct for individual cluster and place it in set RC_i

Step 3: Computation and Ranking of Significant Attributes as follow:

- Computation of descriptor set (I). Find I for cluster C_i, where, where $I = A\text{-}RC_i$ is attribute set and is set of reduct attributes in cluster.
- Computation of PCC score. Calculate PCC for, $dj \in I$: j = 1,..., m, where m is the number of descriptors
- Select the descriptors with PCC threshold >= λ and arrange the descriptors in set I in descending order of CC score

Step 4: Formulate description.

Algo CLUSDESC
Input: I, α, β
Output: P_i
(1) j = 1
(2) $P_i \leftarrow d_j$
(3) while ($I \neq \phi$)
(4) Compute $PE(P_i)$and $CRC(P_i)$
(5) If $PE(P_i) \leq \alpha$ and $CRC(P_i) \geq \beta$ then
(a) if $PE(P_i) \approx 0$ then exit loop;
(6) else; $P_i = P_i \wedge d_{j+1}$: j = 1,..., m //for every conjunctive term dj, it is added only if there is change/decrease in the value of PE.
(7) output P_i
(8) Compute length of P_i

Step 5: Repeat step 2, 3 and 4 for every cluster

EXPERIMENTAL EVALUATION

Dataset characteristics: Proposed RCD approach is tested on four real life datasets of small, medium and large size, from machine learning repository (Blake and Merz, 1998). Characteristics of datasets are presented in Table 1. An attribute can take multiple values from the attribute domain. Hence column Approximate Size of Description Space presents the set of all possible sequence of values an attribute can take in the dataset.

It is not feasible to present the details of experiment on all the datasets due to space constraint. Hence for clarity, detailed steps of proposed approach are presented on two datasets namely Zoo and Iris datasets. However summarize results of soybean and mushroom datasets are also presented.

Experimental results

Zoo dataset: Zoo dataset consist of 101 instances of animals with 16 attributes and 7 output classes. The names of animal constitute the first attribute. 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 attribute. Attribute animal name and class are not considered for clustering.

Table 1: Characteristics of datasets

Dataset	No. of objects	No. of attributes	Approximate size of description space
Zoo	101	16	1×10^5 events
Iris	150	4	1×10^2 events
Soybean disease	47	35	3×10^8 events
Mushroom	8124	22	2×10^{14} events

Table 2: EM Clustering results on Zoo dataset

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

Clustering: EM clustering algorithm with cross validation learnt four clusters from the data instead of seven classes that is known in the dataset. Table 2 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 (Mali and Mitra, 2002).

Reduct computation: Let us consider largest cluster, Cluster 1 for illustration of concepts. Unsupervised reduct computation on Cluster 1 gave two reduct sets R1 {eggs, airborne, aquatic, predator, legs, tail, domestic, catsize} and R2 {airborne, aquatic, predator, toothed, legs, tail, domestic, catsize}. MPCR set which is union of attributes present in the reduct set is considered for removal, as their removal will provide the small set of significant descriptors for ranking. MPCR set resulted in {eggs, airborne, aquatic, predator, toothed, legs, tail, domestic, catsize} from reduct sets R1 and R2. Similarly MPCR is considered for all the clusters. Table 3 shows MPCR set in different clusters.

Computation and ranking of significant attributes: In order to find out set of significant attributes, reduct attributes of Cluster 1 (Table 3) are removed. Removal of reduct attributes has left majority of objects from Cluster1 with similar attribute value pair (descriptor) (Table 4), which are considered significant for generating characteristics of this cluster. Significant descriptors are evaluated on PE (Eq. 1), CRC (Eq. 2) and then PCC (Eq. 3) is computed, which signifies their importance in the cluster. PCC score signify milk = 1 as highest contributing descriptor followed by hair = 1.

Pattern formulation: In order to formulate concise pattern, descriptors with PCC threshold greater than 0.6, is then selected for pattern formulation. Table 5 shows the descriptor set for different clusters along with value of significance in the bracket.

Let us consider pattern formulation for Cluster1. Threshold α of PE less than 0.05 and β of CRC greater than 0.75 is considered for this experiment. This means, if a pattern satisfies 75% objects of a cluster and has PE in the range of 0 to 0.05, then it is a meaningful pattern. Pattern P is assigned first descriptor (milk = 1) and evaluated for CRC and PE. Pattern (milk = 1) supports 41 objects in full dataset that includes all 39 objects of Cluster 1, leading to CRC score 0.97 (Eq. 5) and PE score 0.02 (Eq. 4). In the next step descriptor hair = 1 is concatenated with the previous pattern thereby making new pattern (milk = 1 ^ hair = 1). It resulted in CRC score 0.97 (Eq. 5) and PE score 0.00 (Eq. 4). Pattern generation for Cluster 1 involves conjunction of two descriptors milk and hair, therefore pattern length is two (Eq. 6). Table 6 shows pattern along with value of CRC, PE and L(P) for different clusters.

Table 3: MPCR attributes in individual Clusters

Cluster	Reduct
0	Hair, airborne, predator, toothed, venomous, legs, domestic, backbone, breathes
1	Eggs, airborne, aquatic, predator, toothed, legs, tail, domestic, catsize
2	Airborne, aquatic, predator, domestic, catsize
3	Eggs, milk, aquatic, predator, breathes, venomous, legs, domestic, catsize

Table 4: Significant Descriptors of Cluster 1

Attribute	Attribute value	Support in full dataset	Support in cluster1	PE	CRC	PCC
Hair	0	58	1	0.93	0.03	0.04
	1	43	39	0.07	0.98	0.95
Feathers	0	81	40	0.67	1.00	0.57
Milk	0	60	1	0.97	0.03	0.03
	1	41	39	0.03	0.98	0.97
Backbone	1	83	40	0.70	1.00	0.54
Breathes	1	80	40	0.66	1.00	0.59
Vernomous	0	93	40	0.87	1.00	0.36
Fins	0	84	38	0.75	0.95	0.48
	1	17	2	0.25	0.05	0.19

Table 5: Descriptors with PCC threshold greater than 0.6 in Zoo clusters

Zoo cluster	Values
Cluster 0	
Tail	0(0.93)
Milk	0(0.71)
Catsize	0(0.71)
Eggs	1(0.69)
Cluster 1	
Milk	1(0.97)
Hair	1(0.95)
Cluster 2	
Feathers	1(1)
Legs	2(0.95)
Toothed	0(0.86)
Hair	0(0.72)
Eggs	1(0.72)
Milk	1(0.71)
Cluster 3	
Fins	1(0.85)
Breathes	0(0.79)
Hair	0(0.72)
Toothed	1(0.70)
Milk	0(0.65)

Table 6: Formulated Patterns in different Clusters

Cluster	No. of objects	Pattern	CRC	PE	L(P)
0	20	tail = 0 ^ milk = 0	1	0	2
1	40	milk = 1 ^ hair = 1	0.97	0	2
2	20	feathers = 1	1	0	1
3	20	fins = 1	0.75	0.024	1

Iris dataset: Iris is a well known dataset in pattern recognition literature. This dataset consist of 150 Iris specimens, each measured on four morphological attributes: sepal length, sepal width, petal length and petal width. It is known that there are 50 objects each of three species, Iris-setosa, Iris-versicolor and Iris-virginica. Data is discretized first, as objective of the study is to describe the obtained clusters in terms of attribute and its value range. Class attribute is not considered in clustering.

Clustering: EM clustering algorithm with cross validation learnt three clusters from the data. Table 7 presents EM clustering results on Iris specimens. Cluster 1 contains Iris-setosa specimen as this group is linearly separable from the other two. Cluster 0 and Cluster 2 have majority of specimen from Iris-versicolor and Iris-virginica specimen, respectively.

Reduct computation: In the next stage, unsupervised reduct is computed for individual Iris clusters. Reduct computation for Cluster 2 gave single reduct set R1 = {sepal length, sepal width, petal length, petal width} with all the variables. In this situation, approximate solution for reduct computation is considered (Ref. Subsection Significance of Attribute, under Rough Set Theory), with value of reduct computation function set to 0.8. This iteration gave three reduct sets R1 = {sepal length, sepal width}, R2 = {sepal length, sepal width, petal length} and R3 = {sepal length, sepal width, petal length, petal width}. Then dynamic reduct attributes {sepal length, sepal width, petal length} are considered for removal from Cluster 2. Similarly dynamic reduct attributes are computed for Cluster 0 and Cluster 1. Table 8 shows reduct attributes in different clusters.

Computation and ranking of significant attributes: Let us consider the case of Cluster 0. Reduct attributes of Cluster 0 (Table 8) are removed to find out set of significant descriptors. PCC score is then computed for significant descriptors (petal length and petal width). Table 9 shows different evaluation measures of descriptors from Cluster 0.

Pattern formulation: User specified PCC threshold greater than 0.6 for descriptor selection and threshold α of PE less than 0.05 and β of CRC greater than 0.75 is considered for pattern evaluation in Iris experiment. Table 10 shows the descriptor set for different clusters along with value of significance in the bracket.

Let us consider pattern formulation for Cluster 0. PE value for descriptor petal length in the range (2.45-4.75) is zero. If PE alone is considered as measure of significance, then descriptor petal length has an edge over other descriptors. If the same is considered as pattern of Cluster 0, then this pattern (0.8<petal width<1.75) supports 45 objects in full dataset. This in turns satisfies 44 objects of Cluster 0 leading to CRC(0.84) and PE (0.01). Cluster 0 has support of 53 objects; therefore this pattern is not giving full coverage to objects of this cluster. However, new defined measure of PCC gives an edge to descriptor petal width in the range (0.8-1.75) over petal length. When this is considered as pattern (0.8<petal width<1.75), then it supports 54 objects in full dataset, which in turns satisfies 53 objects of Cluster 0. In the next iteration, conjunction of two descriptors ((0.8<petal width<1.75) ^ (2.45<petal length<4.75)) is considered as pattern, then it supports 44 objects of Cluster 0 (versicolor) without any error. Similarly patterns are formulated for all the clusters. Table 11 shows patterns obtained for different clusters.

Table 7: EM clustering results on Iris dataset

Cluster 0	Cluster 1	Cluster 2
53	50	47

Table 8: Dynamic reduct attributes in Iris clusters

Iris cluster	Attributes
0	Sepal length, sepal width
1	Sepal length, sepal width
2	Sepal length, sepal width, petal length

Table 9: Evaluation measures for descriptors of cluster 0

Attribute	Attribute value	Support in full dataset	Support in cluster 0	PE	CRC	PCC
Petal length	(2.45-4.75)	45	45	0.000	0.849	0.92
	(4.75-5.15)	21	7	0.144	0.132	0.34
	(5.15-6.9)	34	1	0.340	0.018	0.11
Petal width	(0.8-1.75)	54	53	0.010	1.000	0.99

Table 10: Significant descriptors with PCC threshold greater than 0.6 in Iris clusters

Cluster	Petal width	Petal length
0	(0.8-1.75)(0.99)	(2.45- 4.75)(0.92)
1	(0.1-0.8) (1)	(1-2.45) (1)
2	(1.75-2.5) (0.99)	

Table 11: Formulated patterns for Iris clusters

Cluster	No. of objects	Pattern	CRC	PE	L(P)
0 (versicolor)	53	((0.8<petal width<1.75) ^ (2.45<petal length<4.75))	0.83	0.0	2
		(0.8<petal width<1.75)	1	0.01	1
1 (setosa)	50	1≤petal length≤2.45 or 0.1≤petal width≤0.8	1	0.0	1
2 (virginica)	47	1.75≤petal width≤2.5	0.97	0	1

Soybean dataset: Soybean disease dataset contains 47 objects and set of 35 variables characterizing diaporthe-stem-canker, charcoal-rot, rhizoctonia-root-rot and phytophthora-rot diseases. Class variable which specifies the diagnosis of a disease from the set of soybean

diseases is not considered in clustering. All the variables are nominal in nature. It is observed that some variables in the dataset are having same value for all of its instances; hence these variables are considered irrelevant and removed from the dataset. Reduced dataset then has 47 objects and 21 variables characterizing these objects.

EM algorithm with cross validation learnt four clusters from the data. Pattern formulation with proposed approach resulted in concise cluster description giving full coverage to all the cluster objects, without any error.

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

Mushroom dataset: Mushroom dataset is considered bench marking dataset because of its size and attribute domain. Dataset contains description of mushroom samples from three categories (edible, poisonous or not known). Mushrooms under not known category are combined with poisonous category. Dataset consists of large number of 8124 records. The number of edible and poisonous mushrooms in the dataset is 4208 and 3916 respectively. There are 22 categorical variables which describes the physical characteristics of mushrooms. Class variable (edible (e) or poisonous (p)) and variable stalk root with missing values are not considered for clustering.

EM clustering algorithm with cross-validation learned 14 numbers of clusters from the data (Arora *et al*., 2008). Out of 14 clusters, 11 clusters were pure clusters and 3 clusters were mix clusters. Pure clusters in the sense that mushrooms in every cluster are either all poisonous or all edible. Previous study on mushroom dataset by Guha *et al*. (1999), while experimenting with Rock algorithm has also mentioned the segmentation of dataset into pure clusters. As our aim is to generate characteristics of cluster so that specific actions are directed towards individual clusters, hence only pure clusters are considered for pattern formulation for edible and poisonous categories. It is observed that patterns obtained with RCD, distinctively described the clusters with no errors. Patterns obtained are of short length hence easily understandable to users. On average two to three attributes are used to describe the clusters.

CONCLUSION

Clustering provides unsupervised grouping of objects. However the resulting clusters need to be analyzed and understood. Result of RCD approach on datasets is summarized here. It is observed that RCD approach resulted in concise and user understandable patterns. Figure 1 presents the minimum and maximum length of the patterns to describe the clusters as compared to total number of attributes in the dataset. For example, three of the soybean diseases clusters can be described with pattern of length one and one is described using pattern of length two, hence minimum and maximum pattern length is one and two, respectively.

Figure 2 presents the summarized results on coverage and precision error for different datasets. Coverage is computed as the number of objects covered, out of total objects in the dataset using obtained patterns. For example, clustering algorithm on mushroom dataset is has resulted in 780 objects in three mix clusters; hence patterns obtained for pure edible and poisonous clusters covers remaining 7344 objects. Therefore percentage coverage is 90% for mushroom dataset. Correctness is 100-PE and represents the percentage of objects

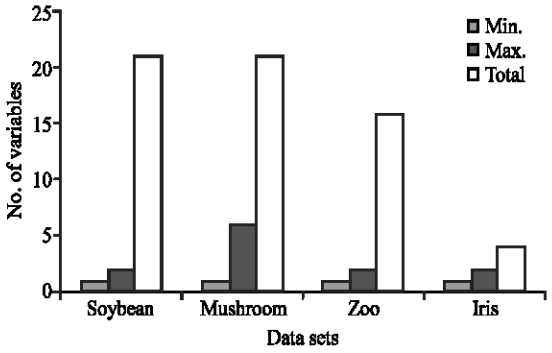

Fig. 1: Number of attributes to describe the clusters

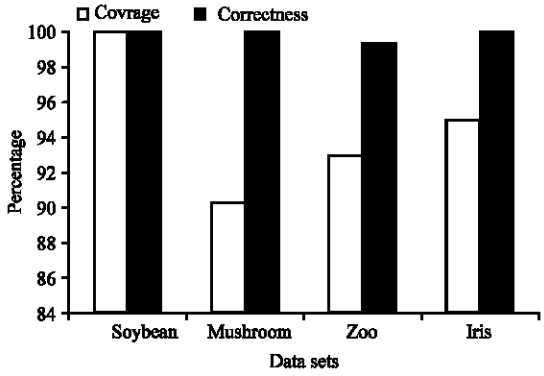

Fig. 2: Average coverage and correctness on datasets

described correctly with proposed approach. Pattern obtained with RCD approach resulted in describing the pure mushroom clusters without any error, hence correctness is 100%. This shows that proposed methods for pattern formulation are promising for large datasets too. Similarly for Zoo dataset average coverage is 93% and average correctness in describing the covered objects is 99%. In case of Iris dataset, pattern formulated with proposed approach is able to cover 95% objects without any error.

It is observed that patterns obtained with proposed approach, distinctively described the clusters with no or minimum errors. Advantage of RCD approach is that, in one pass it removes all the non significant attributes (reduct) from the cluster. In next iteration, it ranks only significant descriptors which are less, as majority of cluster members have similar attribute value pair. Ranking of significant descriptors results in creation of concise cluster description. This approach will trigger future research in the area of cluster description. Future work will be focused on studying the applicability of proposed approach with other clustering algorithms.

REFERENCES

Abidi, S.S.R. and A. Goh, 1998. Applying knowledge discovery to predict infectious disease epidemics. Lecture Notes Comput. Sci., 1531: 170-181-181.

Abidi, S.S.R., M.K. Hoe and A. Goh, 2001. Analyzing data clusters: A rough set approach to extract cluster defining symbolic rules. Lecture Notes Comput. Sci., 2189: 248-257.

Arora, A., S. Upadhyaya and R. Jain, 2008. Learning patterns from clusters using reduct. Proceedings of 6th International Conference on Rough Sets and Current Trends in Computing, Akron, OH, USA., LNCS 5603, Oct. 23-25, Springer Berlin/Heidelberg, pp: 389-398.

Blake, C.L. and C.J. Merz, 1998. UCI Repository of Machine Learning Databases. 1st Edn., University of California, Irvine, CA.

Ganter, B. and R. Wille, 1997. Formal Concept Analysis: Mathematical Foundations. 1st Edn., Springer Verlag, New York Inc., Secaucus, New Jersey, USA., ISBN: 3540627715.

Guha, S., R. Rastogi and K. Shim, 1999. ROCK: A robust clustering algorithm for categorical attributes. Information Systems, pp: 512–521. https://eprints.kfupm.edu.sa/62955/.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann Publisher, San Fransisco, USA., ISBN: 1-55860-901-6.

Holmes, G., A. Donkin and I.H. Witten, 1994. WEKA: a machine learning workbench. Proceedings of the 1994 2nd Australian and New Zealand Conference on Intelligent Information Systems, Nov. 29-Dec. 2, IEEE Computer Society Press, pp: 357-361.

Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data clustering: A review. ACM Comput. Surveys, 31: 264-323.

Mali, K. and S. Mitra, 2002. Clustering of symbolic data and its validation. Lecture Notes Comput. Sci., 2275: 339-344.

Minz, S. and R. Jain, 2005. Refining decision tree classifiers using rough set tools. Int. J. Hybrid Intell. Syst., 2: 133-148.

Mirkin, B., 1999. Concept learning and feature selection based on square-error clustering. Machine Learn., 35: 25-39.

Mirkin, B., 2005. Clustering for Data Mining: Data Recovery Approach. 1st Edn., Chapman and Hall/CRC, USA., ISBN:1584885343.

Öhrn, A., J. Komorowski, A. Skowron and P. Synak, 1998. The Design and Implementation of a Knowledge Discovery Toolkit Based on Rough Sets: The ROSETTA System. In: Rough Sets in Knowledge Discovery 1: Methodology and Applications Polkowski, L. and A. Skowron (Eds.). Physica-Verlag, Heidelberg, USA., pp: 376-399.

Pawlak, Z., J. Grzymala-Busse, R. Slowinski and W. Ziarko, 1995. Rough sets. Commun. ACM, 38: 88-95.

Polkowski, L., 2002. Rough Sets: Mathematical Foundations. 1st Edn., Springer, New York, USA., ISBN: 3790815101.

Skowron, A., J. Komorowski, Z. Pawlak and L. Polkowski, 2002. Rough Sets Perspective on Data and Knowledge: Handbook of Data Mining and Knowledge Discovery. 1st Edn.,. Oxford University Press, Inc. New York, USA, ISBN: 0-19-511831-6, pp: 134-149.

Talavera, L., 1999. Feature selection as retrospective pruning in hierarchical clustering. Lecture Notes Comput. Sci., 1642: 75-86.

Wróblewski, J., 1995. Finding minimal reducts using genetic algorithms. Proceedings of the 2nd Annual Join Conference on Information Sciences, Sept. 28-Oct. 1, Wrightsville Beach, NC., pp: 186-189.