Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 67(1) 2013 33-42

Approach for Mining Multiple Patterns from Clusters

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

Received 9 October 2010; Revised 23 December 2011; Accepted 11 February 2012

SUMMARY

Approach for multiple pattern extraction from obtained individual clusters is presented in this paper. Pattern extraction supports the end users in understanding the cluster concept. Pattern discovery approach utilizes the concept of reduct from rough set theory to find out non-significant attributes in a cluster which has no role in pattern formation. These non-significant attributes (reduct) are removed and remaining attributes are ranked for their significance in the cluster. Multiple pattern formulation approach uses ranked attributes to generate concise cluster patterns. Applicability of the approach is demonstrated using soybean disease and zoo datasets from machine learning repository. Objective of applying proposed approach on soybean disease clusters and clusters of zoo animals is to obtain the patterns to describe those clusters.

Keywords: Clustering, Data mining, Cluster description, Cluster pattern, Multiple patterns, Reduct, Rough set theory.

1. INTRODUCTION

Clustering is an important data mining task, which deals with distribution of data into number of clusters such that data items within a cluster are highly similar, whereas the data items belonging to different clusters are highly dissimilar (Han and Kamber 2001, Holmes *et al*. 1994, Xu and Wunsch 2005). From data mining perspective, the underlying objective of applying clustering technique on dataset is to discover the concept and patterns within the data which can be revealed by grouping the objects into clusters (Mirkin 2005). Wealth of clustering algorithms are available in literature (Han and Kamber 2001, Xu and Wunsch 2005) but majority of them lacks in producing cluster description in the form of pattern. At times post processing of obtained clusters is essential in order to understand the pattern of obtained clusters. Pattern discovery is required in various areas, for example – To develop disease diagnostic system, where there is a need to study the diseases characteristics; In Web Mining: finding pattern in the set of web users; characterization of animal and plant taxonomy; In transcriptome and proteome expression analysis under a set of condition like-Biotic/Abiotic stress, disease/ normal state.

Area of producing cluster pattern for individual cluster is relatively new and there are few approaches in this direction. Mirkin has proposed a method for cluster description and ranking of attributes within cluster based on square error criteria, which is applicable to only continuous attributes (Mirkin 1999). Abidi *et. al*. has proposed the method of pattern creation for unsupervised data using dynamic reduct from Rough Set Theory (RST) (Abidi *et. al*. 2001, Abidi and Goh 1998). They have used the cluster information obtained after cluster finding and generated rules from entire data with respect to decision/cluster attribute. Other popular description approach like decision tree is not directly applicable to clustering as criteria in clustering is to get homogenous clusters with respect to all the attributes (Mirkin 2005). However in decision tree homogeneity is with respect to decision

Corresponding author : Rajni Jain
E-mail address : rajni@ncap.res.in

attribute. Arora *et al.* has proposed the approach for generation of single cluster pattern from obtained clusters with partitional clustering method using reduct from RST (Arora *et al.* 2009, 2008, 2009). These approaches define a cluster with single pattern and it is observed that in certain cases single pattern is insufficient to describe the behavior of a cluster and extra information is required to completely describe the cluster. In this paper, we present generalized approach for pattern formulation in which multiple patterns can be generated for understanding the cluster in a better way. Work carried out in this paper is an extension of the pattern discovery approach given by Arora *et al.* (2009, 2008, 2009) for multiple pattern formulation. For clarity, architecture of pattern discovery approach is presented in section 2 and then multiple pattern formulation method is discussed in section 3. Section 4 provides experimental details of the approach on soybean disease and zoo datasets followed by discussion and conclusion in section 5.

2. ARCHITECTURE OF PATTERN DISCOVERY APPROACH

Functional architecture of pattern discovery given by Arora *et al.* (2009, 2008, 2009) is presented here for clarity of concepts. Approach is broadly categorized into four steps.

Step 1: Partitioning clustering algorithm is applied on dataset in order to obtain non overlapping distinct clusters. 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) (Mirkin 2005, 1999). Given a cluster C_i and a certain value v for attribute a, majority of cluster members will exhibit this value $\{a = v\}; a \in A, v \in V_a\}$ (descriptor). Intuitively, attributes which have similar value for majority of objects in the cluster are considered significant and rest are non significant in generating pattern for that cluster.

Step 2: Post processing of individual clusters is carried out using reduct from RST. RST divides the data into indiscernible/similar classes. Reduct is defined as the minimum set of attributes which can differentiate all indiscernible classes in the dataset. Computation of reduct on individual clusters provides the set of attributes which distinguishes objects in a cluster and is considered non significant in generating pattern for that cluster.

There are different reduct computation algorithms available in literature (Öhrn *et al.* 1998, Polkowski 2002). Genetic Algorithm (GA) (Nguyen and Skowron 2002, Wróbkwski 1995) is selected for reduct computation, as it can produce many reducts of varying cardinality/length. This provides flexibility to the experimenter for selection of attributes from the reduct population produced by GA. There are multiple approaches for selection of attributes from reducts generated by GA, one can refer to literature (Öhrn *et al.* 1998, Polkowski 2002) for the same. We have considered Maximum Possible Combination Reduct (MPCR), which is defined as the union of attributes present in the reducts obtained after applying GA (Jain and Minz 2008, Jain 2004).

MPCR set is considered because it contains set of all attributes which are contributing in distinguishing the objects in the cluster. Moreover removal of attributes in MPCR will lead to comparatively smaller subset of significant descriptors for ranking and pattern formulation.

Step 3: Non significant (reduct) attributes are removed. Significant descriptors for cluster C_i are stored in set I_i. Then in order to generate concise cluster pattern, descriptors $(d_j \in I_i : j = 1, \ldots, m)$ are evaluated on PCC score, which is a real number in the interval [0, 1] and reflects the significance of descriptor in a cluster (Mirkin 1999).

$$PCC(a = v) = \sqrt{(1 - PE(a = v)^{*} CRC(a = v)} \qquad (1)$$

$$PE(a = v) = \frac{Support_U(a = v) - Support_{Ci}(a = v)}{cardU - card\ C_i}$$

$$CRC(a = v) = \frac{card_{Ci}(a = v)}{card\ C_i}$$

$Support\ C_i\ (a = v) = card\ C_i\ (a = v)$ = number of objects satisfying ($a = v$) in cluster C_i.

$Support_U\ (a = v)$ = number of objects satisfying ($a = v$) in universe set.

User can select the descriptors with PCC thresholds, for pattern formulation. Descriptors in set I_i are arranged in decreasing order of PCC score.

Step 4: Pattern P_i of cluster C_i is defined as, $P_i = \overset{m}{\underset{j=1}{\Lambda}} d_j; d_j \in I,$ which is formed by concatenating significant descriptors from cluster C_i in the decreasing order of their PCC score in order to generate concise pattern. Formulated patterns are evaluated on Precision Error (PE) and Coverage ratio in Cluster (CRC) for its accuracy (Nguyen and Skowron 2002).

$$PE(P_i) \quad = \frac{Support_U(P_i) - Support(P_i)}{card\ U - card\ C_i} \tag{2}$$

$$CRC(P_i) \quad = \frac{card_{Ci}(P_i)}{card\ C_i} \tag{3}$$

where $Support_U$ (P_i) and $Support_{Ci}$ (P_i) denote the number of objects satisfying pattern P_i in whole dataset and in cluster C_i respectively. *card U* and *card* C_i denote the number of objects in universe set and cluster C_i respectively. $card_{Ci}$ (P_i) denotes the number of objects in cluster C_i satisfying the pattern P_i.

Pattern Length ($L(P_i)$) is used to measure the conciseness of the pattern and defined as (Nguyen and Skowron 2002):

$$L(P_i) = \textit{Number of descriptors occuring in } P_i \tag{4}$$

In this paper, different steps of pattern discovery approach (Data Clustering, Reduct Computation and Computation & Ranking of Significant Descriptors) are followed in the similar manner (Arora *et al.* 2009, 2008, 2009). Pattern formulation approach has been modified and generalized approach for pattern formulation has been proposed which results in discovery of multiple patterns.

3. MULTIPLE PATTERN FORMULATION APPROACH

There can be large number of hidden patterns in the dataset. Given a set of significant descriptors of size m, there can be $2^m - 1$ possible pattern combinations which can be generated (Öhrn *et al.* 1998). Our aim is not to generate all possible patterns but meaningful and concise patterns. A pattern is considered meaningful and concise, if it has minimal PE, maximum coverage and minimum number of descriptors (Nguyen and Skowron 2002). At times it is not possible to obtain pattern without any error and with full coverage to the cluster. Hence user specified threshold α of PE and β of CRC is input to the algorithm, to generate pattern satisfying those parameters. Steps of the proposed approach are presented here.

Step 1: Input set of significant descriptors ($d_j \in I_i : j = 1, ..., m$) for cluster C_i, which is obtained after applying steps (Data Clustering, Reduct Computation, Computation & Ranking of significant descriptors) of pattern discovery approach. Input threshold for PE (α) and Coverage (β) for pattern evaluation.

Step 2: Consider descriptor from array I_i in descending order of PCC score; ($P_i \leftarrow d_j$); if PCC score is 1, then single descriptor is sufficient to describe the cluster.

Step 3: If PCC score is less than 1 then concatenate the descriptor with next descriptor from array ($P_i = P_i \ \Lambda \ d_{j+1} : j = 1, ..., m$); Every conjunctive term d_j is added to the pattern only if there is decrease in the value of PE; Evaluate the pattern (P_i) on PE and CRC; If PE and CRC are in specified threshold limits then output pattern.

Step 4: Procedure (step 2 and step 3) need to be repeated with every descriptor in the array. Keep on concatenating the descriptors in descending order of their PCC score to obtain multiple pattern satisfying threshold criteria of α and β. Carry out the process of concatenation of descriptors till array is exhausted or desired numbers of patterns are obtained.

4. EXPERIMENTAL DETAILS

Approach is experimented with Soybean disease and Zoo datasets from machine learning repository (Blake and Merz 1998).

4.1. Dataset Description

4.1.1. Soybean disease dataset

Soybean disease dataset consists of 47 objects and set of 35 attributes characterizing four soybean diseases (diaporthe-stem-canker, charcoal-rot, rhizoctonia-root-rot and phytophthora-rot diseases). All the attributes are nominal in nature. This dataset is considered because of its description space. The theoretical size of the description space (i.e, the set of all possible sequences of values of descriptors) is:

$$7 * 2 * 3 * 3 * 2 * ... * 2$$
$$= \text{approx. } 396{,}361{,}728.0$$
$$= 3 * 10^8 \text{ events}$$

It is observed that some attributes in the dataset are having same value for all of its instances, hence data is preprocessed. After preprocessing (removal of irrelevant variables), dataset contains 21 attributes characterizing four soybean diseases. Attribute class information is not considered for clustering.

4.1.2 Zoo dataset

Zoo dataset consist of 101 instances of animals with 17 variables and 7 output classes. The names of animals constitute the first variable. There are 15 boolean variables, with values 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 variable, number of legs {0, 2, 4, 5, 6, 8} correspond to character variable. Thus, the theoretical size of the description space (*i.e.*, the set of all possible sequences of values of descriptors) is:

$$2 * 2 * 2 * \ 2 * ... * 6 = \text{approx. } 196608$$
$$= 1 * 10^5 \text{ events}$$

4.2 Data Clustering

As our aim is to study the pattern of clusters, hence partitional algorithm is selected as it partitions dataset into distinct non overlapping clusters. There exist large numbers of partitioning clustering algorithms in literature (Blake and Merz, Holmes *et al*. 1994, Mirkin 2005). The choice of clustering algorithm depends both on the data types and application. Expectation Maximization (EM) algorithm is popular partitioning clustering algorithm which can handle both nominal and continuous attributes hence it has been selected for this experimentation.

EM algorithm assigns a probability distribution to each object which indicates the probability of it belonging to each of the clusters (Han and Kamber 2001, Holmes *et al*. 1994, Mirkin 2005). EM algorithm assumes data is from multivariate Gaussian distribution and sets the parameter of the Gaussian to maximize the log likelihood of the data. The algorithm is characterized as (i) Initialize k cluster centers, (ii) Iterate between two steps.

- **Expectation step (E):** calculate cluster probability for each instance.
- **Maximization step (M):** estimate distribution parameters from cluster probabilities to maximize the likelihood of instances given the specified number of clusters.

EM algorithm estimates the k (number of clusters) present in the data using cross validation method. EM algorithm determines the k from the data by maximizing the logarithm of the likelihood of future data and the same is estimated using ten fold cross-validation (Holmes *et al*. 1994). Cross validation process begins with one cluster and it continues to add clusters until the estimated log-likelihood decreases which determines the optimal number of cluster present in the data.

Weka (Holmes *et al*. 1994) is open source data mining software which is used for clustering objects using EM algorithm.

EM clustering algorithm on soybean disease dataset, learnt four clusters from the dataset with cross validation. Four disease clusters corresponding to diseases *diaporthe-stem-canker* (Cluster1), *charcoal-rot* (Cluster2), *rhizoctonia-root-rot* (Cluster3), *phytophthora-rot* (Cluster4) are obtained. Structure of obtained clusters is presented in Table 1.

Table 1. EM clustering results on soybean disease dataset

Cluster Name	No. of objects
Cluster 1	10
Cluster 2	10
Cluster 3	10
Cluster 4	17

EM clustering algorithm on Zoo dataset, learnt four clusters from the data with cross validation instead of seven classes that is known in the dataset. Table 2 presents 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 (Arora *et al.* 2009).

Table 2. EM clustering results on Zoo dataset

Cluster Name	No. of objects
Cluster 1	21
Cluster 2	40
Cluster 3	20
Cluster 4	20

4.3. Reduct Computation

To study the disease characteristics, reduct analysis is carried out individually on obtained clusters. Rosetta is an open source software which is used to compute reduct based on GA (Öhrn *et al*. 1998, Wróblewski 1995).

Let us consider the case of soybean Cluster 1 for *diaporthe-stem-canker* disease. Reduct computation for Cluster1 produced 17 reducts of cardinality/length 3 to 5 attributes. Table 3 present reducts for Cluster 1.

Table 3. Reducts obtained using GA for Cluster 1

R1 = {v1, v6, v7}	R2 = {v1, v6, v20}
R3 = {v1, v8, v10}	R4 = {v1, v5, v10, v20}
R5 = {v5, v6, v9, v20}	R6 = {v5, v8, v9, v10}
R7 = {v5, v6, v9, v10}	R8 = {v1, v5, v9, v10}
R9 = {v5, v6, v7, v8}	R10 = {v1, v5, v7, v10}
R11 = {v5, v6, v8, v10}	R12 = {v1, v6, v9, v10}
R13 = {v5, v6, v8, v20, v22}	R14 = {v5, v6, v10, v20, v22}
R15 = {v5, v6, v7, v10, v22}	R16 = {v5, v6, v7, v9, v22}
R17 = {v5, v7, v8, v10 v22}	

MPCR set, which is the union of attributes present in reducts, {v1, v5, v6, v7, v8, v9, v10, v20, v22} is computed from reduct population produced by GA as given in Table 3. Similarly reduct computation and MPCR set computation is carried out for individual soybean disease clusters. Table 4 presents the MPCR attributes for different soybean disease clusters.

Similarly, to study the characteristics of animal clusters from Zoo dataset, reduct analysis is carried out individually on obtained four clusters. Table 5 presents the MPCR attributes in different animal clusters.

Table 4. MPCR attributes in soybean disease clusters

Cluster	MPCR attributes
Cluster 1	v1, v5, v6, v7, v8, v9, v10, v20, v22
Cluster 2	v1, v4, v5, v6, v7, v9, v10, v20
Cluster 3	v1, v2, v5, v6, v8, v9, v10, v12, v20, v25, v35
Cluster4	v1, v3, v4, v5, v6, v8, v9, v10, v21, v24

Table 5. MPCR attributes in Zoo clusters

Cluster	Reduct
Cluster1	hair, airborne, predator, toothed, venomous, legs, domestic, backbone, breathes
Cluster2	eggs, airborne, aquatic, predator, toothed, legs, tail, domestic, catsize
Cluster3	airborne, aquatic, predator, domestic, catsize
Cluster4	eggs, milk, aquatic, predator, breathes, venomous, legs, domestic, catsize

4.4 Ranking of Descriptors on Precision Coverage Coefficient

In the next step for pattern discovery, MPCR attributes obtained for individual clusters are removed and PCC score (Eq. 1) is computed for rest of the significant descriptors in the corresponding clusters.

Table 6. Significant descriptors for soybean clusters with PCC score greater than 0.5

Cluster1	v21 = 3(1), v23 = 1(1), v28 = 0(0.85), v2 = 0(0.82), v4 = 1(0.78), v24 = 1(0.75), v35 = 0(0.69), v3 = 2(0.61), v26 = 0(0.51), v27 = 0(0.51)
Cluster2	v3 = 0(1), v21 = 0(1), v22 = 3(1), v26 = (1), v27 = 1(1), v28 = 0(0.27), v24 = 0(0.83), v2 = 0(0.82), v8 = 1(0.71), v35 = 0(0.69), v23 = 0(0.51)
Cluster3	v22 = 1(0.91), v4 = 0(0.90), v21 = 1(0.88), v24 = 1(0.75), v28 = 3(0.73), v2 = 1(0.65), v7 = 1(0.69), v3 = 2(0.61), v23 = 0(0.51), v26 = 0(0.51), v27 = 0(0.51)
Cluster4	v22 = 2(1), v35 = 1(0.98), v2 = 1(0.85), v28 = 3(0.81), v7 = 1(0.73), v23 = 0(0.57), v26 = 0(0.57), v27 = 0(0.57), v12 = 1(0.54), v20 = 0(0.51)

Table 6 presents the significant descriptors for different soybean disease clusters ranked on PCC score and arranged in descending order of PCC score. PCC score for descriptors are given in parenthesis.

In case of animal clusters from Zoo dataset, descriptors with PCC threshold greater than 0.60 is considered for pattern formulation. Table 7 shows the descriptor set for different clusters along with value of significance in the parenthesis.

Table 7. Descriptors with PCC threshold greater than 0.6 in Zoo clusters

Cluster1	tail = 0(0.93), milk = 0(0.71),
	catsize = 0(0.71), eggs = 1(0.69)

Cluster2	milk = 1(0.97), hair = 1(0.95)

Cluster3	feathers = 1(1), legs = 2(0.95),
	toothed = 0(0.86), hair = 0(0.72)

4.5. Multiple Pattern Formulation

Let us consider the case of Cluster1 of soybean disease for illustration of multiple pattern formulation approach. Threshold for CRC and PE is considered 1 and 0 respectively for pattern evaluation. This means, only those pattern are considered meaningful, which gives complete coverage (CRC score 1) without any error (PE score 0).

- Consider the first descriptor v21 = 3 (Table 6) with highest PCC score as pattern. As the PCC score is 1 for this descriptor, therefore single descriptor is sufficient to describe the cluster completely. Pattern (v21 = 3) evaluation for descriptor with PCC score 1 will always give complete coverage (CRC 1) without any error (PE 0).
- In order to formulate, another pattern, next descriptor in the array (v23 = 1) is considered as pattern. Again descriptor (v23 = 1) has PCC score 1, therefore single descriptor is sufficient to describe the cluster.
- For third pattern, next descriptor (v28 = 0) is considered as pattern, and evaluated for PE and CRC. On evaluation, this gives PE value 0.27 (Eq. 2) and CRC value 1 (Eq. 3).
    - PE score exceeds the specified threshold, hence next descriptor (v2 = 0) is concatenated and new pattern (v28 = 0 ^ v2 = 0) is evaluated. There is no change in the value of PE with the concatenation of descriptor (v2 = 0). As per proposed approach, every conjunctive term is added to the pattern only if there is decrease in the value of PE, hence v2 = 0 is not concatenated and previous pattern (v28 = 0) is retained.
    - Next descriptor (v4 = 1) is concatenated with previous pattern; and pattern (v28 = 0 ^ v4 = 1) is evaluated. This gives PE value 0.11 and CRC value 1; again PE score exceeds the specified threshold.
    - Next descriptor (v24 = 1) is concatenated and pattern (v28 = 0 ^ v4 = 1 ^ v24 = 1) is evaluated; which gives CRC score 1 and PE 0. Hence this is a meaningful pattern satisfying the threshold of CRC and PE and therefore retained in the output. Then no further concatenation of descriptors is carried out in this order.
- For fourth pattern, descriptor (v28 = 0) is concatenated with next descriptor from array

Table 8. Multiple patterns for Cluster1 (diaphore-stem-canker)

Pattern	CRC	PE	L(P)

v21 = 3
stem-cankers = above-sec-nde
	1	0	1

v23 = 1
fruiting-bodies = present
	1	0	1

v28 = 0 ^ v4 = 1 ^ v24 = 1
fruit-pods = norm temp = norm ^
externaldecay = firm-and-dry
	1	0	3

v28 = 0 ^ v24 = 1
fruit-pods = norm ^ external decay
= firm-and-dry
	1	0	2

v28 = 0 ^ v3 = 2
fruit-pods = norm ^ precip = gt-norm
	1	0	2

v28 = 0 ^ v26 = 0
fruit-pods = norm ^ int_discolor = none
	1	0	2

v28 = 0 ^ v27 = 0
fruit-pods = norm ^ sclerotia = absent
	1	0	2

v2 = 0 ^ v4 = 1 ^ v24 = 1
plant_stand = normal ^ temp = norm
external decay = firm-and-dry
	1	0	3

v4 = 1 v24 = 1 ^ v35 = 0
temp = norm ^ external decay
= firm-and-dry roots = norm
	1	0	3

v4 = 1 ^ v35 = 0 ^ v3 = 2
temp = norm ^ roots = norm ^
precip = gt-norm
	1	0	3

(v24 = 1) and new pattern (v28 = 0 ^ v24 = 1) is evaluated. This gives complete coverage to the cluster without any error.

Similarly, this process is continued to generate patterns till all the selected descriptors are exhausted or desired number of patterns are obtained. Table 8 presents the patterns for Cluster1 (diaphore-stem-canker).

Process is repeated for all the soybean disease clusters. Multiple patterns for remaining disease clusters are shown in Annexure A. Obtained patterns are analyzed on pattern length for conciseness. Table 9 presents the summarized results on conciseness of patterns for different disease clusters. Entries in Table 9, presents number of patterns with different pattern lengths in individual clusters. Let us consider for example Cluster1. There are total 10 patterns obtained with complete coverage and without any error. There are 2 patterns of pattern length 1, 4 patterns of pattern length 2 and 4 patterns of pattern length 3.

Table 9. Number of different length patterns for soybean disease clusters

	Pattern Length L(P)			
	1	2	3	4
Cluster 1	2	4	4	–
Cluster 2	5	4	3	–
Cluster 3	–	5	2	–
Cluster 4	1	2	1	2
Total	**8**	**15**	**10**	**2**

On computing the average length of pattern in different clusters, it is observed that on an average 2 descriptors are involved in pattern formulation

$$\frac{1*8+2*15+3*10+4*2}{8+15+10+2} \approx 2$$

In case of animal clusters, we have specified threshold "α" of PE less than 0.05 and β of CRC greater than 0.75 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. Multiple patterns discovered for Zoo clusters using proposed approach are shown in Annexure B. Let us consider the patterns formulated for Cluster3 (Table 10), which clearly indicate that this is a cluster of birds with feathers. Other supporting characteristics are that, they have two legs and do not have tooth; they have two legs and do not have hair; they have two legs and lay eggs and they have two legs and do not give milk.

Table 10. Patterns for Cluster3 (Zoo dataset)

Pattern	CRC	PE	L(P)
feathers = 1	1	0	1
legs = 2 ^ toothed = 0	1	0	2
legs = 2 ^ hair = 0	1	0	2
legs = 2 ^ eggs = 1	1	0	2
legs = 2 ^ milk = 0	1	0	2

Table 11 summarizes the number of patterns from Zoo clusters, obtained with threshold 0.75 and 0.05 for CRC and PE respectively from Annexure B. Patterns are formulated using significant descriptors with PCC threshold greater than 0.6. Entries in Table 11 present the number of patterns. For example, there are 5 patterns which give complete coverage to Zoo clusters without any error (Annexure B). Similarly, there are 3

Table 11. Number of patterns from Zoo clusters

PE CRC	0.0	0.01	0.02	0.03
1.00	5	–	–	1
0.95	3	–	–	1
0.90	1	1	–	2
0.85	–	–	–	–
0.80	–	–	–	–
0.75	1	–	2	–

Table 12. Number of different length patterns for Zoo clusters

	Pattern Length L(P)		
	1	2	3
Cluster 1	–	3	1
Cluster 2	1	1	–
Cluster 3	1	4	–
Cluster 4	1	4	1
Total	**3**	**12**	**2**

patterns which give 95% coverage, 1 pattern with 90% coverage and 1 pattern with 75% coverage and all these patterns describe the clusters without any error (first column of Table 11).

Table 12 presents the summarized results on conciseness for Zoo clusters from Annexure B. Entries in Table 12, presents number of patterns with different pattern length in individual clusters. On computing the average, it is found that on an average 2 descriptors are involve in pattern formulation.

5. DISCUSSION AND CONCLUSION

On summarizing the results for soybean disease clusters, it is observed that multiple patterns are obtained for soybean disease clusters which provided complete coverage to clusters without any error. On an average, 2 descriptors are involved in pattern formulation for disease clusters. In case of Zoo clusters, patterns are obtained which give 75% coverage to cluster objects and error is in 0.05% range. On an average, 2 descriptors are involved in pattern formulation for different Zoo clusters. Hence Patterns obtained with proposed approach are meaningful. Further, obtained patterns are concise and easily understandable by the user. In future, multiple pattern formulation approach will be experimented with other real world datasets from different domains to study the effectiveness of the approach in generating cluster pattern.

REFERENCES

Arora, A., Upadhyaya, S. and Jain, R. (2009). Integrated approach of reduct and clustering for mining patterns from clusters. *Inform. Tech. J.*, **8(2)**, 173-180.

Arora, A., Upadhyaya, S. and Jain, R. (2008). Learning Patterns from Clusters using Reduct, Rough Sets and Current Trends in Computing*, Lecture Notes in Computer Science*, **5306/1**, 389-398. Springer Berlin, New York.

Arora, A., Upadhyaya, S. and Jain, R. (2009). Post processing of clusters for pattern discovery: Rough set approach. *J. Ind. Soc. Agril. Statist.*, **63(2)**, 181-188.

Abidi, S.S.R., Hoe, K.M. and Goh, A. (2001). Analyzing data clusters: A rough set approach to extract cluster defining symbolic rules, Lecture notes in Computer Science. *Springer Verlag: Berlin*, **2189**, 248-257.

Abidi, S.S.R. and Goh, A. (1998). Applying knowledge discovery to predict infectious disease epidemics. Lecture Notes in *Computer Science*. *Springer Verlag: Berlin*, **1531**, 170-181.

Blake, C. and Merz, C. (yyyy). UCI repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences, http://www.ics.uci.edu/~mlearn/.

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

Holmes, G., Donkin, A. and Witten, I.H. (1994). WEKA: a machine learning workbench. In proc. of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, 357-361, Software, available at: http://www.cs.waikato.ac.nz/~ml/

Jain, R. and Minz, S. (2008). Drawing conclusions from forest cover type data- The hybridized rough set model. *J. Ind. Soc. Agril. Statist.*, **62(1)**, 75-84.

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*, Berlin: Springer-Verlag, 3-98.

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

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

Öhrn, A., J. Komorowski, Skowron, A. and Synak, P. (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, edited by L. Polkowski, and A. Skowron, 376-399 Heidelberg: Physica-Verlag, Software available at http://www.rossata.com.

Polkowski, L. and Rough Sets (2002). *Mathematical Foundations*. Springer.

Nguyen, S.H. and Skowron, A. (2002). Searching for relational pattern on data. In: J. Komorowski, J. Zytkow(eds.), *Proceedings of First European Symposium on Principals of Data Mining and Knowledge Discovery-PKDD'97, Trondheim, Norway, LNAI 1263, Berlin*, Springer Verlag, 265-276.

Wróblewski, J. (1995). Finding minimal reducts using genetic algorithms. In pror. of the Second Annual Join Conference on Information Sciences, 186-189 Wrightsville Beach, NC. Also in ICS Research report 16/95, Warsaw University of Technology.

Xu, R. and Wunsch, D. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, **16(3)**, 645-978.

Annexure A

Multiple Patterns for Soybean Clusters

Multiple patterns for Cluster 2 (charcoal-rot)

Pattern	CRC	PE	L(P)
Precip = lt-norm	1	0	1
stem-cankers = absent	1	0	1
canker-lesion = tan	1	0	1
int-discolor = black	1	0	1
sclerotia = present	1	0	1
fruit-pods = norm ^ external decay = absent	1	0	2
fruit-pods = norm ^ severity = pot-severe ^fruiting bodies = absent	1	0	3
fruit-pods = norm ^ fruiting bodies = absent	1	0	2
external decay = absent ^ plant_stand = normal	1	0	2
external decay = absent ^ severity = pot severe ^ roots = norm	1	0	3
external decay = absent ^ roots = norm	1	0	2
plant_stand = normal ^ severity = pot-severe ^ ruiting bodies = absent	1	0	3

Multiple Patterns for Cluster 3 (rhizoctonia-root-rot)

Pattern	CRC	PE	L(P)
canker-lesion = brown ^ temp = lt-norm	1	0	2
canker-lesion = brown ^ stem-cankers = below-soil	1	0	2
canker-lesion = brown ^ fruit-pods = dna	1	0	2
canker-lesion = brown ^ area-damaged = low-areas ^ fruiting-bodies = absent	1	0	3 `
canker-lesion = brown ^ fruiting-bodies = absent	1	0	2
temp = lt-norm ^ stem-cankers = below-soil ^ external decay = firm-and-dry	1	0	3
temp = lt-norm ^ external decay = firm-and-dry	1	0	2

Multiple Patterns for Cluster 4 (phytophthora-rot)

Pattern	CRC	PE	L(P)
canker-lesion = dk-brown-blk	1	0	1
root = abnorm ^ leaves = abnorm	1	0	2
Plant_stand = normal ^ leaves = abnorm	1	0	2
fruit-pods = dna ^ leaves = abnorm ^ lodging = yes	1	0	3
fruiting-bodies = absent ^ int-discolor = none ^ leaves = abnorm ^lodging = yes	1	0	4
fruiting-bodies = absent ^ scleroti a = absent ^ leaves = abnorm ^ lodging = yes	1	0	4

Annexure B

Multiple Patterns for Zoo Clusters

Multiple patterns for Cluster 1

Pattern	CRC	PE	L(P)
tail = 0 ^ milk = 0	0.95	0.000	2
tail = 0 ^ catsize = 0	0.90	0.012	2
tail = 0 ^ catsize = 0 ^ eggs = 1	0.90	0.000	3
tail = 0 ^ eggs = 1	0.95	0.000	2

Multiple Patterns for Cluster 2

Pattern	CRC	PE	L(P)
milk = 1	0.97	0.032	1
milk = 1 ^ hair = 1	0.97	0.000	2

Multiple Patterns for Cluster 3

Pattern	CRC	PE	L(P)
Feathers = 1	1	0	1
legs = 2 ^ toothed = 0	1	0	2
legs = 2 ^ hair = 0	1	0	2
legs = 2 ^ eggs = 1	1	0	2
legs = 2 ^ milk = 0	1	0	2

Multiple Patterns for Cluster 4

Pattern	CRC	PE	L(P)
fins = 1	0.75	0.024	1
fins = 1 ^ hair = 0	0.75	0.000	2
fins = 1 ^ toothed = 1	0.75	0.024	2
hair = 0 ^ toothed = 1	1.00	0.037	2
hair = 0 ^ toothed = 1 ^ milk = 0	0.90	0.037	3
Toothed = 1 ^ milk = 0	0.90	0.037	2