International Journal of Computational Intelligence Systems, Vol.2, No. 1 (March, 2009), 10-16

REDUCT DRIVEN PATTERN EXTRACTION FROM CLUSTERS

SHUCHITA UPADHYAYA
Deptt. of Computer Science and Application, Kurukshetra University
Kurukshetra-136119, India.
shuchita_bhasin@yahoo.com

ALKA ARORA
Indian Agricultural Statistics Research Institute
Library Avenue, Pusa, New Delhi-110012, India
alkak@iasri.res.in, alka27@yahoo.com

RAJNI JAIN
National Center for Agricultural Economics and Policy Research
Library Avenue, Pusa, New Delhi-110012, India
rajni@ncap.res.in

Received: 27-05-2008
Revised: 10-10-2008

Abstract

Clustering algorithms give general description of clusters, listing number of clusters and member entities in those clusters. However, these algorithms lack in generating cluster description in the form of pattern. From data mining perspective, pattern learning from clusters is as important as cluster finding. In the proposed approach, reduct derived from rough set theory is employed for pattern formulation. Further, reduct are the set of attributes which distinguishes the entities in a homogenous cluster, hence these can be clear cut removed from the same. Remaining attributes are then ranked for their contribution in the cluster. Pattern is formulated with the conjunction of most contributing attributes such that pattern distinctively describes the cluster with minimum error.

Keywords: Clustering, Cluster description, Data mining, Knowledge discovery, Pattern, Rough set theory, Reduct

1. Introduction

Fast developing computer science and engineering techniques has made the information easy to capture process and store in databases. Discovery of knowledge from this huge amount of data is a challenge indeed. Knowledge discovery in databases (KDD), popularly known as data mining is an attempt to make sense of the information embedded in large databases[1]. Clustering is a key area in data mining. The underlying assumption of clustering in data mining is to find out the hidden patterns in data, which can be revealed by grouping the entities into clusters. Clustering algorithms partitions a given dataset into clusters such that entities in a cluster are more similar to each other than entities in different clusters. Clustering algorithms in literature are broadly classified into hierarchical and partitional methods (See Refs. 1, 2, 3 and 4 for details on different clustering algorithms). Hierarchical algorithms construct a tree like structure (dendogram) combining all the entities. Description is subjective in case of dendogram. Partitional method divides the entities into k non overlapping clusters, where k is the number of clusters specified by the user as input. K-means and Expectation Maximization (EM) algorithms are the widely known partitional algorithms. These clustering algorithms only generate general description of the clusters depicting number of clusters and member entities of each cluster. However, it lacks in generation of underlying pattern in the dataset, as this approach has no mechanism for selecting and evaluating the attributes in the process of

Copyright: the authors

generating clusters[5]. Hence post processing of clusters is required for deriving useful knowledge in the form of pattern. Pattern is formulated by conjunction of significant attribute value pair and hence describes the cluster in more meaningful format.

Cluster description is useful in studying the relationship that describes the underlying data. It also helps in interpretation of clusters in user understandable format. Cluster description can be applied in various areas for understanding the patterns, viz., disease diagnostic system (to study the disease characteristics), Web mining (to find pattern in the set of Web users), tourism industry (to find what features of places and tourist attract each other) and banks (to identify defaulters).

Rough Set Theory (RST) proposed by Pawlak[6], 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 equivalence/indiscernible classes; each indiscernible class can be considered as natural cluster. Moreover, RST performs automatic concept approximation by producing minimal subset of attributes called reduct, which can distinguish all the indiscernible classes in the dataset.

Proposed approach is applied as post processing step after obtaining clusters using partional clustering algorithm. Reduct is computed for individual clusters as our aim is to formulate pattern of individual clusters. Reduct is the set of attributes which distinguishes the entities in a homogenous cluster, hence can be clear cut removed from the same. It is anticipated that remaining attributes will then have similar attribute value pair. These remaining attributes play significant role in pattern formulation. Remaining attributes are ranked for their contribution in the cluster, and then pattern is formulated with the conjunction of major contributing attributes.

The paper is organized as follows: rough set concepts are discussed in section 2. In section 3, basic notion of cluster description along with related work in the area of cluster description is provided. In section 4, proposed reduct driven cluster description approach is discussed. In section 5, steps of proposed approach are demonstrated on hypothetical animal dataset along with results of proposed approach on real life datasets from standard UCI repository. In section 6, conclusion is provided.

2. Rough Set Theory Concepts

In RST, data is represented as an information system $X = (U, A \bigcup \{d\})$,(See Refs. 6, 7, 8 and 9 for details on RST). In this, U is non-empty finite set of entities and A is a non-empty, finite set of attributes on U, where $d \notin A$ is decision/class attribute. With every attribute $a \in A$, we associate a set V_a such that $a : U \to V_a$. The set V_a is called the domain or value set of attribute a. Every entity x, in the information system X, is characterized by its information vector:

$$InfX(x) = \{(a, a(x)) : a \in A\} \quad (1)$$

Relationship between entities is described by their attribute values. Indiscernibility relation $IND(B)$, for any subset $B \subseteq A$ is defined by:

$$x\ IND(B)\ y \Leftrightarrow \forall_{a \in B}(a(x) = a(y)) \quad (2)$$

That is, two entities 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. Entities in the information system about which we have the 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 is the set of attributes which can differentiate all equivalence classes. Johnson and Genetic Algorithm (GA) are popular reduct computation algorithms[10]. Johnson algorithm produces single reduct set of minimal cardinality however genetic algorithm produces many reduct sets of varying cardinality/length. Reduct set may contain more than one attribute. Common attributes shared by all the reduct sets, produced by GA[11] is called Core. However Maximum Possible Combined Reduct (MPCR) [12] set is defined as the union of attributes present in the reduct sets obtained by GA.

3. Cluster Description

3.1. *Cluster Description Concepts*

In the information system, attribute value pair of the form $\{(a = v); a \in A, v \in V_a\}$ is defined as descriptor D[7]. Let $(a = v)_U$ and $(a = v)_C$ denote the set of entities satisfying $(a = v)$ in whole dataset and in cluster C respectively.

Copyright: the authors

Hence $\sup port_U(a=v) = card([a=v]_U)$ and $\sup port_C(a=v) = card([a=v]_C)$.

Pattern P of cluster C is defined as, $P = (D_1 \wedge D_2 \wedge D_3 \wedge ..... \wedge Dn)$, which is formed by concatenating significant descriptors from cluster C.

It is quite possible that some entities that do not belongs to C, also satisfies P. Therefore as defined by Mirkin[13], pattern is evaluated on Precision Error (PE). PE of pattern P for cluster C denoted by PE (P) is the number of entities that lie outside cluster C, for which pattern P is true divided by number of entities outside cluster C.

$$PE(P) = \frac{\left|\sup port_U(P) - \sup port_C(P)\right|}{\left|card\ U - card\ C\right|} \quad (3)$$

3.2. *Review of Literature*

The field of producing cluster description for individual clusters is relatively new. There are few references of cluster description approaches available in literature. Mirkin[13] has proposed a method for cluster description applicable to only continuous attributes. In this, 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 in the sorted order. Forward attribute selection process stops only after the last element of attribute set is checked. Abidi et. al.[14, 15] has proposed the rough set theory based method for rule creation for unsupervised data using dynamic reduct. Dynamic reduct is the frequently occurring reduct in the population of reduct sets obtained using genetic algorithm. However these approaches have limitations. Mirkin[13] approach is applicable only to datasets having continuous attributes. Abidi et. al.[14, 15] has used the cluster information obtained after cluster finding and generated rules from entire data with respect to decision attribute, instead of producing description for individual clusters. Other popular description approach like decision tree[2] is not directly applicable to clustering as criteria in clustering is to get homogenous clusters with respect to all the attributes. However in decision tree homogeneity is with respect to decision attribute.

4. Reduct driven Cluster Description (RCD) Approach

The proposed Reduct driven Cluster Description (RCD) approach is applicable to clusters found by the clustering algorithm. Proposed approach of pattern formulation is divided into three stages. First stage deals with obtaining clusters from dataset by applying clustering algorithm. In the second stage we have computed sets of significant and non significant attributes for that cluster. Cluster is set of similar data entities, therefore attributes which has similar value for majority of its attribute value pair are considered significant for that cluster and rest are non significant. As reduct accounts for discerning between the entities, therefore computation of reduct set (RC) in a cluster will provide the set of non significant attributes for that cluster. These non significant attributes (reduct) can be straight forward removed from the cluster.

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 class/decision attribute. Such reducts are referred as unsupervised reducts in this paper. In the present study, we have computed unsupervised reduct for individual clusters in comparison to reduct computation for dataset, as our aim is to generate patterns of individual clusters.

Removal of reduct attributes from the cluster provides set of significant attributes (I) for that cluster. An attribute value pair (descriptor) is said to be highly significant, if all the entities satisfying that descriptor belongs to a single cluster. It is quite possible that some entities those satisfy the descriptor also belongs to other clusters. Therefore we proposed to evaluate descriptors for their significance on Precision Error (PE), which is defined as:

$$PE(D) = \frac{\left|\sup port_U(a=v) - \sup port_C(a=v)\right|}{\left|card\ U - card\ C\right|} \quad (4)$$

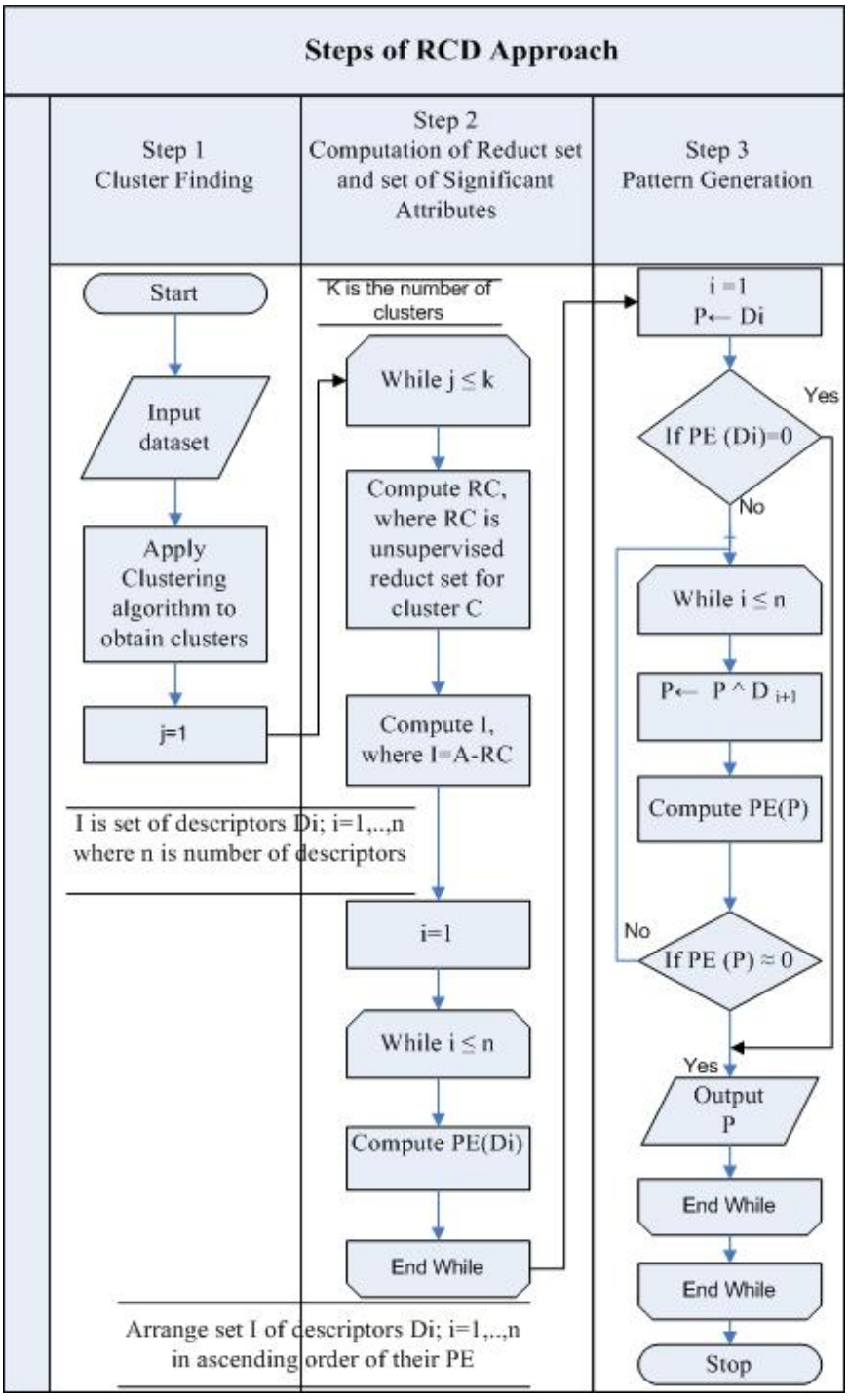
PE is calculated for every descriptor (D) in the set of significant descriptors (I). Descriptors in set I are then arranged in ascending order of their PE. If a descriptor has zero PE, which means all the entities satisfying that descriptor belongs to a single cluster.

In the third stage method is proposed for formulation of pattern P. P is formulated by combining

Copyright: the authors

the descriptors with less PE such that PE (P) is minimum, hence *P* distinctively describes the cluster with minimum or without errors. Step2 and Step 3 needs to be carried out for every cluster.

4.1. ***Flow Chart of RCD Approach:***



Copyright: the authors

5. Examples of Application

5.1. *Data Description*

Steps of RCD approach is illustrated using small dataset from animal taxonomy[16] (Table 1). Dataset consists of information on phylogenetic characteristics of animals from amphibians and mammals classes. The classes however are not informed.

Concepts of RST introduced in section 2 are explained here using Table 1. Universal set U= {frog, toad, elephant, cat, dog, rabbit, jaguar, whale} contains eight animal entities. These entities are characterized by the attribute set A= {Metabolism, Cover, Dentition, Reproduction, #Feet}. Domain sets of different attributes are:

$V_{Metabolism}$ = {ectothermic, endothermic},
V_{Cover} = {wetskin, hair},
$V_{Dentition}$ = {superior, no, complete, the_youngest},
$V_{Reproduction}$ = {oviparous, viviparous} and
$V_{\#Feet}$ = {4, 0}.

Entities are characterized by the attribute value, therefore Information vector InfX(frog)= {Metabolism, ectothermic}, i.e value of attribute Metabolism for entity frog is ectothermic.

For any subset $B \subseteq A$, when B={Metabolism}then entities (frog, toad) (elephant, cat, dog, rabbit, jaguar, whale) in these sets are indiscernible and form equivalence classes.

Hence, U/IND(B)={(frog, toad) (elephant, cat, dog, rabbit, jaguar, whale)}. Similarly when B= {Metabolism, Dentition} then U/IND(B)= {(frog) (toad) (elephant, cat, dog, rabbit, jaguar) (whale)}.

Table1: Animal dataset

Name	Metabolism	Cover	Dentition	Reproduction	#Feet
frog	ectothermic	wetskin	superior	oviparous	4
toad	ectothermic	wetskin	no	oviparous	4
elephant	endothermic	hair	complete	viviparous	4
cat	endothermic	hair	complete	viviparous	4
dog	endothermic	hair	complete	viviparous	4
rabbit	endothermic	hair	complete	viviparous	4
jaguar	endothermic	hair	complete	viviparous	4
whale	endothermic	hair	the_youngest	viviparous	0

5.2. *Data Clustering*

We have used Weka[17] implementation of EM algorithm for cluster finding. EM models the distribution of entities probabilistically, so that an entity belongs to a cluster with certain probability[2]. The first step, calculation of the cluster probabilities, which are the expected class value, is “expectation”; the second step calculation of the distribution parameter is “maximization” of the likelihood of the distribution given the data. Weka implementation of EM algorithm can handle different types of attributes, and it 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.

EM algorithm partitioned the animal dataset into two clusters. Cluster1 has entities of elephant, cat, dog, rabbit, jaguar and whale, therefore forming the cluster of mammals. Cluster2 has entities of frog and toad; therefore forming the cluster of amphibians.

5.3. *Reduct Computation*

We have used Rosetta[10] implementation of genetic algorithm for computation of unsupervised reduct. To study the characteristics of amphibians and mammals, reduct analysis is carried out on these individual clusters. Reduct analysis on Cluster1 has resulted in two reduct sets, R1= {Dentition} and R2= {# Feet}. Reduct accounts for distinguishing the entities in the cluster; therefore we have considered all possible reduct attributes in the form of MPCR set, hence MPCR = {Dentition, #Feet}. Reduct computation on Cluster2 gave single reduct set, R3= {Dentition}.

Copyright: the authors

5.4. *Cluster Description*

Cluster description provides value addition to obtained clusters by describing the cluster structure in terms of relevant attributes. Let us consider Cluster 1(mammals) for pattern formulation. Reduct attributes (Dentition, #Feet) of Cluster1 are removed from the same. Cluster1 is then left with the set of descriptors (Metabolism=endothermic, Cover=hair, and Reproduction=viviparous) having similar value for all the entities. We then calculated PE for these descriptors to find out the major contributing descriptors. Let us consider computation of PE (Equ. 4) for descriptor (Metabolism=endothermic) from Cluster1. Descriptor Metabolism=endothermic has support of 6 entities in the full dataset, and Cluster1 has support of all the 6 entities.
Hence, PE (Metabolism=endothermic) = (6-6)/ (8-6) = 0
Similarly PE (# Feet=4)= (7-2)/(8-2)=5/6 for Cluster 2.

Table 2 presents the significant descriptors along with value of PE in bracket for different clusters.

Table 2: Significant descriptors in clusters

Cluster1	Metabolism=endothermic(0), Cover=hair(0), Reproduction=viviparous(0)
Cluster2	Metabolism=ectothermic(0), Cover=wet(0), Reproduction=oviparous(0), #Feet=4(5/6)

Table2 shows that descriptors, Metabolism=endothermic or Cover=hair or Reproduction=viviparous have zero PE, hence either of these descriptors can describe the Cluster1 without any error. Similarly, removal of reduct attribute (Dentition) has left the set of significant descriptors (Metabolism=ectothermic, Cover=wet, Reproduction=oviparous, and #Feet=4) for Cluster2. Table 2 shows that descriptors, Metabolism =ectothermic, Cover=wet, Reproduction=oviparous have zero PE, hence either of these descriptors can describe the Cluster2 without error.

5.5. *Results on Real Life Datasets*

Three datasets of agriculture domain are considered from UCI[18] repository for evaluation of RCD approach. Table 3 shows the characteristics of datasets along with number of clusters obtained by applying EM algorithm. These datasets have unique and class attributes which are not considered for clustering.

Table 3: Characteristics of datasets

Dataset Name	# of entities	# of attributes	# of clusters
Soybean Disease	47	35	4
Zoo	101	17	4
Mushroom	8124	22	14

Objective of applying the RCD approach on soybean disease dataset is to study the disease characteristics of obtained four diseases clusters. Table 4 presents the disease characteristics obtained using RCD approach on this dataset.

Table 4: Soybean disease characteristics

Cluster	# of entities	Pattern	PE
Cluster1 (diaporthe-stem-canker)	10	stem-cankers= above-sec-nde or fruiting-bodies= present	0
Cluster2 (charcoal-rot)	10	precip= lt-norm or stem-cankers= absent or canker-lesion=tan or int-discolor=black or sclerotia=present	0
Cluster3 (rhizoctonia-root-rot)	10	canker-lesion=brown ^ temp= lt-norm	0
Cluster4 (phytophthora-rot)	17	canker-lesion= dk-brown-blk	0

Objective of applying the RCD approach on Zoo dataset is to characterize the animal clusters. Table 5 presents the results on Zoo dataset.

Table 5: Characteristics of animal clusters

Cluster	# of entities	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	.024

RCD approach is evaluated on large benchmarking mushroom dataset. Objective of applying the RCD approach on Mushroom dataset is to study the characteristics of edible and poisonous mushrooms[19].

Copyright: the authors

Table 5 presents the characteristics of mushroom clusters.

Table 5: Mushroom clusters pattern

Cluster	# of entities	Pattern	PE
Poisonous Clusters			
Cluster1	288	spore-print-color=h ^ odor=f ^ cap-surface=s	0
Cluster2	1728	gill-color=b	0
Cluster8	256	odor=p	0
Cluster9	1296	ring-type=l	0
Edible Clusters			
Cluster4	192	stalk-color-above-ring=o or stalk-color-below-ring=o	0
Cluster5	768	gill-spacing =w ^ habitat=g ^ ring-type=e	0
Cluster6	96	gill-spacing=w ^ gill-size=n ^ habitat=d ^ bruises=t	0
Cluster7	1728	habitat=d ^ bruises=t ^ odor=n	0
Cluster10	511	bruises=t ^ stalk-shape=e ^ ring-type=p ^ stalk-surface-below-ring=y ^ gill-size=b ^ ring-number=o.	0
Cluster12	192	stalk-surface-below-ring=y^ cap-surface=y ^ bruises=t	0
Cluster14	288	ring-number=t ^ gill-spacing=w	0

6. Conclusion

Clustering provides unsupervised grouping of objects. However the resulting clusters need to be analyzed and understood. In this paper, RCD approach is presented for selection of significant attributes from individual clusters. Reduct and Maximum Possible Combination Reduct (MPCR) set significantly contributed in removal of non significant attributes from the cluster. Reduct along with Precision Error has resulted in formulation of concise and user understandable patterns from the clusters. It is observed that pattern obtained with RCD, distinctively described the clusters with no or minimum errors. This approach will trigger future research in the area of cluster description.

References

1. B. Mirkin, *Clustering for Data Mining: Data Recovery Approach* (Chapman & Hall/CRC, 2005).
2. J. Han and M. Kamber, *Data Mining: Concepts and Techniques* (Morgan Kaufmann, 2006).
3. A.K. Jain, M.N. Murty and P.J. Flynn, Data Clustering: A review, *ACM Computing Surveys*, 31(3) (2001) 264-323.
4. R. Xu and D. Wunsch II, Survey of Clustering Algorithms, *IEEE Transaction on Neural Networks*, 16(3)(2005) 645-678.
5. R.S. Michalski and R.E. Stepp, Clustering, *Encyclopedia of Artificial Intelligence*, eds. S.C. Shapiro (J. Wiley & Sons, New York, 1992).
6. Z. Pawlak, J. Grzymala-Busse, R. Slowinski, W. Ziarko, Rough Sets, *Communications of the ACM*, 38(1995), 11, 88-95.
7. S. H. Nguyen, T. T. Nguyen, A. Skowron and P. Synak, Knowledge discovery by Rough Set Methods, in *Proc. of the International Conference on Information Systems Analysis and Synthesis* (Orlando USA,1996) 26-33.
8. L. Polkowski, Rough Sets: Mathematical Foundations (Springer, 2002).
9. A.Skowron, J. Komorowski, Z. Pawlak, L. Polkowski, Rough sets perspective on data and knowledge, *Handbook of data mining and knowledge discovery*, 134-149(Oxford University Press, Inc. New York, 2002)
10. Rosetta, http://www.rosetta.com
11. J. Wroblewski, Finding minimal reducts using genetic algorithms, in *Wang (513, 1995)* 186-189.
12. R. Jain, Rough Set based Decision Tree Induction for Data Mining, *Ph. D. Thesis*, JNU, New Delhi 110067, INDIA (2004).
13. B. Mirkin, Concept Learning and Feature Selection based on Square-Error Clustering, *Machine Learning*, 35(1999) 25-40.
14. S. S. R Abidi and A. Goh, Applying knowledge discovery to predict infectious disease epidemics, in H. Lee and H. Motoda (eds.) *Lecture notes in artificial intelligence*, PRICAI'98 (Springer Verlag, Berlin, 1998).
15. S. S. R Abidi, K. M. Hoe and A. Goh, Analyzing data clusters: A rough set approach to extract cluster defining symbolic rules, in H. Fisher and A. Hoffman (eds.) Lecture notes in Computer Science: Advances in Intelligent Data Analysis, 4th Intl. Symposium, IDA-01 (Springer Verlag, Berlin, 2001).
16. H. A. d. Prado, P. M. Engel and H.C. Filho, Rough clustering: An alternative to find meaningful cluster by using the reducts from a dataset, Lecture Notes in Computer Science, 234-238 (Springer, Berlin, 2002)
17. WEKA: A Machine Learning Software, http://www.cs.waikato.ac.nz/~ml/.
18. UCI: Repository of machine learning databases, http://www.ics.uci.edu/~mlearn/.
19. S. Upadhyaya, A. Arora, R. Jain, Deriving Cluster Knowledge Using Rough Set Theory, Journal of Theoratical and Applied Information Technology, 4(8) (2008) 688-696.

Copyright: the authors