J. Basic. Appl. Sci. Res., 2(6)5915-5920, 2012
© 2012, TextRoad Publication
ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

Clustering Using Rough-Set Feature Selection.

Dr Usman Qamar[1], Prof John A. Keane[2]

[1]Computer Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad Pakistan
[2]School of Computer Science, University of Manchester, Manchester, M13 9PL

ABSTRACT

`Feature selection aims to remove features unnecessary to the target concept. Rough-set theory (RST) eliminates unimportant or irrelevant features, thus generating a smaller (than the original) set of attributes with the same, or close to, classificatory power. Clustering, also a form of data grouping, groups a set of data such that intra-cluster similarity is maximized and inter-cluster similarity is minimized. As with classification, there exists a group of attributes or features on the basis of which clustering is carried out; hence RST may be used for clustering.
This paper analyses the effects of rough sets on clustering using 10 datasets, each including a decision attribute. This generates a framework for applying rough-sets for clustering purposes. Rough-sets are then used for knowledge discovery in clustering and the conclusion indicated a very significant result that removal of individual numeric attributes has far more effect on clustering accuracy than removal of categorical attributes.
KEYWORDS: Rough-sets, Classification, Clustering, Feature Selection, Categorical and Numerical Data.

1. INTRODUCTION

Feature selection techniques aims to reduce the number of unnecessary, irrelevant, or unimportant features [1]. It is common practice to use a measure to decide the importance and necessity of features. Rough-set theory (RST) is an extension of set theory for the study of systems characterized by insufficient and incomplete information [2].

RST was proposed by Pawlak [2-3] for knowledge discovery in datasets. Not all attributes in an information system may be required and thus they can be eliminated without losing essential information. Rough-sets provide a method to determine for a given information system the most important attributes in terms of classification accuracy. The concept of the *reduct* is fundamental in RST. A reduct is the essential part of an information system (related to a subset of attributes) which can discern all objects discernible by the original set of attributes of an information system.

Classification is an example of machine learning. Rough-sets have been applied for classification in various applications and have been proved to be useful [3]. Clustering, a form of data grouping is a discovery process that groups a set of data such that the intra-cluster similarity is maximized and the inter-cluster similarity is minimized [4-5]. Just as with classification, there exists a group of attributes or features on the basis of which clustering is carried out. This suggests that RST might be useful for clustering applications

This has given rise to the concept of rough clustering [4–6]. Lingras and West [7] provided rough k-means algorithm [8, 9]. The aim of which was to use rough sets as a k-means clustering. The rough k-means [7] along with its and its extensions [10–11] have been found to be effective in a various clustering applications. The focus of the above papers was on analysing the success or otherwise of rough sets in terms of clustering. The focus here is on not only analysing the success or otherwise of rough sets but moving towards an understanding of why. This will allow developing a framework of how rough-sets may be applied for clustering.

The rest of this paper is structured as follows: Section 2 provides background to RST; the experiment design is discussed in Section 3; analysis of the results is given in Section 4; with conclusions given in Section 5.

2. Rough Set Theory

RST determines the degree of attributes dependency and their significance.

An information system (IS) ([2]) is a representation of a flat table. An IS ($\mathcal{A}$) consists of a pair (U,A), where U is a non-empty, finite set of objects and A is a non-empty, finite set of attributes [2].

***Corresponding Author:** Dr Usman Qamar, Computer Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad Pakistan. Email: usmanq@ceme.nust.edu.pk

$$A = (U, A)$$

Decision systems (DS) [2] are a special kind of IS. By labeling the objects of A, it is possible to construct classes of objects. These classes can then be modeled using rough set analysis. The labels are the target attribute of which to obtain knowledge

IS $A = (U,A)$ and $B \subseteq A$ then we can approximate decision class X using the information contained by the attribute set of B. Thus allows us to define the lower and upper approximations as [3]:

$$X : \underline{B}X = \{x | [x]_B \subseteq X\}$$

$$X : \overline{B}X = \{x | [x]_B \cap X \neq 0\}$$

The difference between the upper and the lower approximation is the the boundary region [2] which can be defined as

$$X : BN_B(X) = \overline{B}X - \underline{B}X$$

Computing reducts is a non trivial task that cannot be solved by a simple increase of computational resources. It is, in fact, one of the bottlenecks of the rough set methodology [2]. Fortunately, there exist good heuristics based on genetic algorithms that compute sufficiently many reducts in often acceptable time [2].

3. Experimentation

Datasets from various different fields are selected such as banking, medicine and census data; datasets selected must posses certain specific properties including different categorical and numerical attributes.

In total 10 datasets from various fields are used for experimentation. Table 1 gives details of the datasets selected.

Table 1: Dataset descriptions

Dataset No	Dataset Name	No of Categorical Attributes	No of Numeric Attributes	Decision Attribute	Source
D1	SARS 2001 (a) Census Data	10	12	Rare/ Not Rare	[14]
D2	SARS 2001 (b)Census Data	12	9	Rare/ Not Rare	[14]
D3	Credit card Approval	8	6	credit card accepted/rejected	[15]
D4	Heart Disease (a)	7	7	Diagnosed with heart disease or not	[16]
D5	Heart Disease (b)	8	10	Diagnosed with heart disease or not	[16]
D6	Wisconsin Breast Cancer (a)	6	10	iagnosed with breast disease or not	[17]
D7	Income Dataset	8	9	Individual's income is above or below $50,000	[18]
D8	Wisconsin Breast Cancer (b)	7	5	Diagnosed with breast disease or not	[17]
D9	Housing prices	7	10	Housing prices is in top 20% or bottom 80%	[18]
D10	HSV patients	10	7	diagnosed or not	[18]

For generating clusters gCLUTO [13] is used. It provides a wide-range of clustering algorithms that operate either directly on the original feature-based representation of the objects or on the object-to-object similarity graphs.

The experiment consisted of repeating the following steps for each dataset.

1. Clustering using complete set of attributes: Using gCluto, the dataset is clustered with the complete set of attributes into 5 way-clustering. The clustering algorithm used is the Bi-Section Method [13].
2. Reduce: The GA is used to compute the reducts for the dataset. Reducts are generated for the dataset using Rosetta.
3. Clustering using Reducts: The dataset is again clustered using the same parameters as used in step 1 but instead of using all the attributes this time only those attributes which are reducts are used. Thus for each reduct, the dataset is clustered each time. For example, if during step two, five sets of reducts were generated, the dataset will be clustered five times, each time with a different set of reducts.
4. Comparison: The clusters of step1 and step 3 are compared. Comparison is done such that for each cluster generated in step 1 and step 3, similarity among the records present in each cluster is noted. The average percentage similarity for the five clusters is called "percentage similarity".

4. RESULTS AND ANALYSIS

Clustering is a form of data grouping, which may benefit from rough-set feature selection. This hypothesises that reducts should produce clustering patterns similar to those produced by the full set of attributes. Table 2 shows the percentage similarity between the original cluster and the reduct cluster for each dataset S. For clustering, the maximum percentage similarity between the original and the reduct cluster is 86% while the minimum similarity is 6%.

Table 2: Percentage Similarity between original and reduct cluster

Dataset	Minimum percentage similarity between the reduct cluster and the original cluster	Maximum percentage similarity between the reduct cluster and the original cluster
D1	21	71
D2	14	77
D3	11	76
D4	6	81
D5	9	77
D6	14	86
D7	18	76
D8	12	77
***D9**	16	84
D10	16	82

Other observations obtained from the clustering results are:

- With clustering, 15% of the reducts produced 80% or more similarity between the original cluster and the reduct cluster; while for classification this was reversed i.e 85% of the reducts generated produced classification error of less than 20%.
- The size of the reduct has an influence on the similarity percentage. Similarity between the original cluster and the reduct cluster increases as the size of the reduct increases. This is shown in Table 3.

Table 3: Percentage similarity distribution

Size of reducts	% Similarity
3	30 to 40
4	40 to 70
5	70 to 100

Reducts provide the attributes that are keys to maintaining minimum loss of clustering. This implies that these attributes are more significant than the attributes that are not part of the reducts. Table 4 shows the instances of numerical and categorical attributes in the reducts generated by a genetic algorithm using Rosetta.

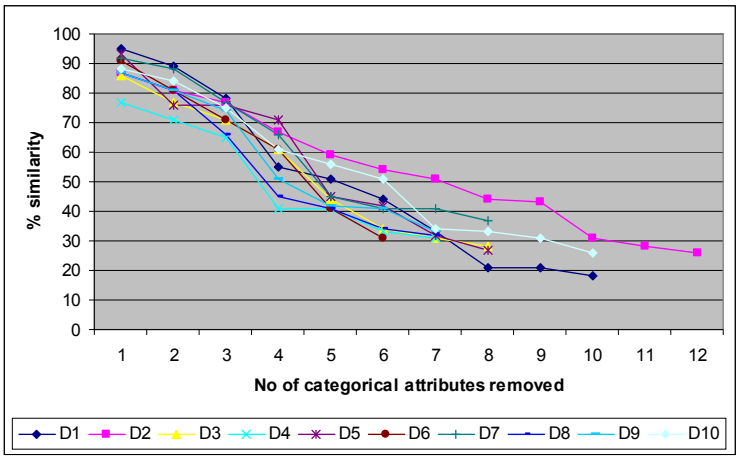
Table 4: Numerical and Categorical attributes in reduct generation.

Dataset	Total no of reducts	Total no of attributes in all reducts	Total % of Numerical attributes in all reducts	Total % of Categorical attributes in all reducts
D1	31	114	56%	44%
D2	28	104	61%	39%
D3	18	71	64%	36%
D4	24	94	61%	39%
D5	17	66	55%	45%
D6	22	102	58%	42%
D7	19	82	63%	37%
D8	18	75	61%	39%
D9	25	101	62%	38%
D10	24	121	62%	38%

As there are more numerical than categorical attributes in the reducts this may suggest that numerical attributes are of greater significance to clustering.

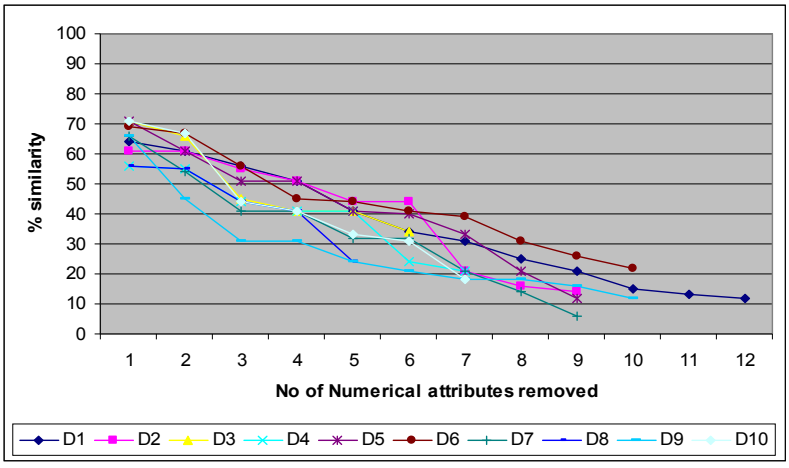
Graph 1 show how the average percentage similarity varies between the original cluster generated by the full set of attributes and the cluster generated by removing an increasing number of categorical attributes from the complete set of attributes.

Graph 1: Effect on % similarity by removing categorical attributes



Graph 2 shows how the average percentage similarity varies between the original cluster generated by the full set of attributes and the cluster generated by removing an increasing number of numerical attributes from the complete set of attributes.

Graph 2: Effect on % similarity by removing numerical attributes



As can be seen from graph 1, removing 30% of categorical attributes from each dataset results in a percentage similarity of between 70% and 80%. In comparison, by removing just a single numerical attribute the percentage similarity falls to 60%. When 30% of the categorical attributes have been removed from each dataset the percentage similarity falls between 55% and 40%. Graphs 1 and 2 suggest that numerical attributes tend to have more significance than categorical attributes in clustering.

This indicates that numerical attributes have a much stronger influence on clustering than as compared to categorical. This can be explained by the fact that categorical attributes may only have limited range of values, e.g. for a categorical attribute such as "Marriage Status" the range of values can be "Single, Married, Divorced, Widow". However for a numerical attribute such as "Years of Work Experience" the possibilities are far more. This means that numerical attributes may hold more information than categorical attributes and as clustering is done on the basis of this information, removal of numerical attributes have a more direct influence on clustering then categorical attributes.

5. Conclusion

This paper analyses the effects of using rough-sets on clustering. The resulting accuracy is considered and mapped to the type and number of attributes both in the original and the reduced datasets.

This paper not only shows the effectiveness of the rough-set feature selection for clustering of data but also provides a general framework for applying rough sets for classification.

The important points are:

- Rough-set feature selection becomes more effective as the number of attributes of the original dataset is increased. One explanation of this result is that the larger the set of original features the more likely it contains redundant or irrelevant features and thus more effective the rough-set will be [2].
- The number of rules and the depth of decision tree generated using the rough-sets is less than the number of rules/depth of decision tree generated using the original dataset. The fewer the rules, the more quickly data will be clustered as indicated in rough sets based PCM [12].
- The size of reduct has an influence on clustering accuracy. Greater the size of the reduct, more the accuracy of clustering.

It also shows that individual numerical attributes have a greater influence on the clustering.

6. REFERENCES

[1] Liu H. and Motoda H. 1999. Feature Selection for Knowledge Discovery and Data Mining. Kluwer: Academic Publishers.

[2] Komorowski J., Pawlak Z. and Skowron. 1999. Rough Sets: A tutorial. World Scientific Publishing Co.

[3] Ohrn A. 1999. Rough Sets: A Knowledge Discovery Technique for Multifactor Medical Outcomes.

[4] Hirano, S., Tsumoto, S. 2000. Rough clustering and its application to medicine. J. Inf. Sci. 124, 125–137.

[5] Peters, J.F., Skowron, A., Suraj, Z., Rzasa, W., Borkowski, M. 2000. Clustering: a rough set approach to constructing information granules. In: Proceedings of 6th International Conference on Soft Computing and Distributed Processing, pp. 57–61.

[6] Voges, K.E., Pope, N.K., Brown, M.R. 2003. A rough cluster analysis of shopping orientation data. In: Proceedings Australian and New ZealandMarketing Academy Conference, Adelaide, pp. 1625–1631.

[7] Lingras, P., West, C. 2004. Interval set clustering of web users with rough k-means. J. Intell. Inf. Syst. 23(1), 5–16.

[8] Hartigan, J.A.,Wong, M.A. 1979. Algorithm as136: a k-means clustering algorithm. Appl. Stat. 28, 100–108.

[9] MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297.

[10] Peters, G. 2006. Some refinements of rough k-means. Pattern Recognition vol. 39, 1481–1491.

[11] Mitra, S. 2004. An evolutionary rough partitive clustering. Pattern Recognition. Lett. 25(12), 1439–1449.

[12] Kaiiali, M.; Wankar, R.; Rao, C.R.; Agarwal, A. 2010. A Rough Set based PCM for authorizing grid resources. 10th International Conference on Intelligent Systems Design and Applications (ISDA), pp- 391 – 396.

[13] Karypis G.: Cluto a clustering toolkit, http://www.cs.umn.edu/cluto.

[14] The Samples of Anonymised Records (SARS), http://www.ccsr.ac.uk/sars/

[15] Quinlan, J. R.: C4.5: Programs for machine learning, Morgan Kaufmann, www.cse.unsw.edu.au/~quinlan/.

[16] Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

[17] UCI Knowledge Discovery in Datasets Archive, http://kdd.ics.uci.edu/databases/census-income/census-income.html

[18] Wisconsin Breast Cancer datasets http://research.cmis.csiro.au/rohanb/outliers/breat-cancer/