(CCQ) Clustering based on Classification Quality

Iwan Tri Riyadi Yanto[1], Rd Rohmat Saedudin[2], Dedy Hartama[3], Tutut Tutut Herawan[4]

[1]Department of Information System, Ahmad Dahlan University, Indonesia
yanto.itr@is.uad.ac.id
[2]Department of Industrial Engineering, Telkom University,Bandung, West Java, Indonesia
rdrohmat@telkomuniversity.ac.id
[3]Departement of Information System, Tunas Bangsa AMIK and STIKOM, Indonesia
dedyhartama@amiktunasbangsa.ac.id,
[4]Department of Information System, University of Malaya
tutut@um.edu.my

Abstract. Clustering a set of objects into homogeneous classes is a fundamental operation in data mining. Categorical data clustering based on rough set theory has been an active research area in the field of machine learning. However, pure rough set theory is not well suited for analyzing noisy information systems. In this paper, an alternative technique for categorical data clustering using Variable Precision Rough Set model is proposed. It is based on the classification quality of Variable Precision Rough theory. The technique is implemented in MATLAB. Experimental results on three benchmark UCI datasets indicate that the technique can be successfully used to analyze grouped categorical data because it produces better clustering results.

Keywords : *Clustering; Rough set; Variable precision rough set model, classification quality*

1. Introduction

Cluster analysis is a data analysis tool used to group data with similar characteristics. It has been used in data mining tasks such as unsupervised classification and data summation, as well as segmentation of large heterogeneous data sets into smaller homogeneous subsets that can be easily managed, separately modeled and analyzed [1]. The basic objective in cluster analysis is to discover natural groupings of objects [2].

A variety of clustering algorithms exists to group objects having similar characteristics. But the implementations of many of those algorithms are challenging in the process of dealing with categorical data. While some of the algorithms cannot handle categorical data, others are unable to handle uncertainty within categorical data in nature [3]. Several clustering analysis techniques for categorical data exist to divide similar objects into groups. Some are able to handle uncertainty in the clustering process, whereas others have stability issues [4].

Recently, many attentions have been put on categorical data clustering, where data objects are made up of non-numerical attributes. For categorical data clustering, a new trend has become in algorithms which can handle un-certainty in the clustering process. One of the well-known techniques is based on rough set theory [5, 6, 7]. The first attempt on using rough set theory for selecting a clustering (partitioning) attribute was proposed by Mazlack et al. [8]. Mazlack proposed a technique called TR (Total Roughness) which is based on accuracy of approximation of a set [5], where the highest value is the best selection of attribute. One of the successful pioneering rough clustering for categorical data techniques is Minimum-Minimum Roughness (MMR) proposed by Parmar et al. [9]. The algorithm for selecting a clustering attribute is based on the opposite of accuracy of approximation of a set [5]. To this, TR and MMR possibly provide the same result on selecting a clustering attribute. The algorithms are based on lower and upper approximations of a set [5,6,7]. However, the original rough set model is quite sensitive to noisy data [10] and some limitation was reported in [11]. There are drawbacks, particularly losing more useful information for demanding the inclusion of the absolutely precision in the classical definition of rough set. In order to overcome the drawback, Ziarko [11] proposed the VPRS model to deal with noisy data and uncertain information by introducing an error parameter β, where $0 \le \beta < 0.5$ as a new way to deal with the noisy data.

Inspired VPRS for handling noisy data, in this paper, we propose an alternative technique for categorical data clustering that there are addresses above issue. For selecting the clustering attribute, it is based on the classification quality of Variable Precision Rough theory.

2. Variable Precision Rough Set

Variable precision rough set (VPRS) extends rough set theory by the relaxation of the subset operator [11]. It was proposed to analyze and identify data patterns which represent statistical trends rather than functional. The main idea of VPRS is to allow objects to be classified with an error smaller than a certain pre-defined level. This introduced threshold relaxes the rough set notion of requiring no information outside the dataset itself.

Definition 1. Let a set U as a universe and $X,Y \subseteq U$, where $X,Y \neq \phi$. The error classification rate of X relative to Y is denoted by $e(X,Y)$, is defined by

$$e(X,Y)=\begin{cases}1-\dfrac{|X\cap Y|}{|X|} & ,\ |X|>0\\ 0 & ,\ |X|=0\end{cases}. \tag{1}$$

Definiton 2. Let U be a finite set and a set $X \subseteq U$. Given β be a real number within the range $0 \le \beta < 0.5$. The B_β -lower approximation of *X*, denoted by $\underline{B}_\beta(X)$ and B_β -upper approximation of *X,* denoted by $\overline{B}_\beta(X)$, respectively, and are defined by

$$\underline{B}_\beta(X)=\{x\in U: e([x]_B, X)\le \beta\}\ \overline{B}_\beta(X)=\{x\in U: e([x]_B, X)<1-\beta\}. \tag{2}$$

The set $\underline{B}_\beta(X)$ is called the positive region of *X*. It's the set of object of *U* that can be classified into X with error classification rate not greater than β . Then we have $\underline{B}_\beta(X)\subseteq\overline{B}_\beta(X)$ if only if $0 \le \beta < 0.5$, which means that β be restricted in an interval $[0,0.5)$ in order to keep the meaning of the "upper" and "lower" approximations. The attributes dependency degree of rough set model in Variable Precision Rough Set Model is called the measure of classification quality. Based on Ziarko's notions, it is given in the following definition.

Definition 3. The accuracy of approximation variable precision (accuracy of variable precision roughness) of any subset $X \subseteq U$ with respect to $B \subseteq A$ is denoted by $\alpha_{B_\beta}(X)$. It is presented as

$$\alpha_{B_\beta}(X)=\frac{\left|\underline{B}_\beta(X)\right|}{\left|\overline{B}_\beta(X)\right|} \tag{3}$$

where $|X|$ denotes cardinality of X. If $\beta = 0$, it is the traditional rough set model of Pawlak.

Proposition 4. Let $S=(U,A,V,f)$ be an information system, $\alpha_B(X)$ be an accuracy of roughness and $\alpha_{B_\beta}(X)$ is an accuracy of variable precision roughness given β the error factor of variable precision. $(0 \le \beta < 0.5) \Rightarrow \alpha_B(X) \le \alpha_{B_\beta}(X)$.

Proof. Based on Definition 5, if $\beta \ge 0.5$, then $\underline{B}_{\beta}(X) \not\subset \overline{B}_{\beta}(X)$. Thus, for $0 \le \beta < 0.5$, we have $\underline{B}_{0}(X) \supseteq \underline{B}_{\beta}(X)$ and $\overline{B}_{0}(X) \subseteq \overline{B}_{\beta}(X)$. Conse-quently $|\underline{B}_{0}(X)| \le |\underline{B}_{\beta}(X)|$ and $|\overline{B}_{0}(X)| \ge |\overline{B}_{\beta}(X)|$.

For $\beta = 0$, based on Definition 5, $\alpha_{B}(X) = \alpha_{B_{\beta}}(X)$.

For $0 < \beta < 0.5$, we have $|\underline{B}(X)| \le |\underline{B}_{\beta}(X)|$ and $|\overline{B}_{\beta}(X)| \le |\overline{B}(X)|$. Hence

$$\frac{|\underline{B}(X)|}{|\overline{B}(X)|} \le \frac{|\underline{B}_{\beta}(X)|}{|\overline{B}_{\beta}(X)|}.$$

Therefore, $\alpha_{B}(X) \le \alpha_{B_{\beta}}(X)$. □

Definition 5. Let *S = (U, A,V, f)* be an information system and let *D* and *C* be any subsets of *A*. Given β be a real number within the range $0 \le \beta < 0.5$. The measure of classification quality of attribute C on attributes D, denoted by $D \Rightarrow_{\gamma} C$, is defined by

$$\gamma = \frac{\sum_{X \in U/D} |\underline{C}_{\beta}(X)|}{|U|}, \tag{4}$$

Obviously, $0 \le \gamma \le 1$. Attribute *D* is depends on *C* with the classification error not greater than β if elements of the universe *U* can be classified to equivalence classes of the partition *U/D*, employing *C*.

3. Classification quality for selecting clustering attribute

In this section, we will present the proposed technique, which is clustering based on classification quality of Variable Precision Rough Set (CCQ). The technique uses the classification quality in variable precision of attributes of rough set theory.

Proposition 9. Let *S = (U, A,V, f)* be an information system and let *D* and *C* be any subsets of *A*. Given β be a real number within the range $0 \le \beta < 0.5$ .if *D* depends on *C* with the classification error not greater than β, then $\alpha_{D_{\beta}}(X) \le \alpha_{C_{\beta}}(X)$, for every $X \subseteq U$.

Proof. Let *D* and *C* be any subsets of *A* in information system. From the hypothesis, we have the portioning *U/D* is finer that *U/C*. Therefore, for every $x \in X \subseteq U$, $e([X]_C, X) \le e([X]_D, X)$. And hence, for every $X \subseteq U$, we have

$$\underline{D}_\beta(X) \subseteq \underline{C}_\beta(X) \subseteq \overline{C}_\beta(X) \subseteq \overline{D}_\beta(X).$$

Consequently

$$\alpha_{D_\beta}(X) = \frac{\left|\underline{D}_\beta(X)\right|}{\left|\overline{D}_\beta(X)\right|} \le \frac{\left|\underline{C}_\beta(X)\right|}{\left|\overline{C}_\beta(X)\right|} = \alpha_{C_\beta}(X).$$

□

The attribute with highest average of classification quality is selected as the clustering decision.

Definition 10. Suppose $a_i \in A$, $V(a_i)$ has k-different values, say y_k, $k = 1,2,\cdots,n$. Let $X(a_i = y_k)$, $k = 1,2,\cdots,n$ be a subset of the objects having *k*-different values of attribute a_i. The measure of classification quality of the set $X(a_i = y_k)$, $k = 1,2,\cdots,n$ for given β error factor, with respect to a_j, where $i \ne j$, can be generalized as follows

$$\gamma_{a_i}(a_j) = \frac{|X_{\beta_{a_j}}(a_i = y_1)|}{|U|} + \frac{|X_{\beta_{a_j}}(a_i = y_2)|}{|U|} + \cdots + \frac{|X_{\beta_{a_j}}(a_i = y_n)|}{|U|}$$

$$\gamma_{a_i}(a_j) = \frac{\sum_{U/a_j} |X_\beta(a_i = y_k)}{|U|}|, \quad k = 1,2,\cdots,n. \tag{5}$$

Definition 11. Given n attributes, mean classification quality of attribute $a_j \in A$ with respect to $a_j \in A$, where $i \ne j$, denoted as $CCQ(a_i)$ is obtained by following formula

$$CCQ(a_i) = mean\left(\gamma_{a_i}(a_j)\right), i, j < n. \tag{6}$$

3.1. Example

The following table is a Discretized of supplier information system containing 15 objects with 4 categorical-valued conditional attributes; Demand Delivery WH, Production Plan, Sales forecast, and Supply. Then, we will select a clustering attribute among all candidates.

Table 1. A Discretized Supplier information system

	D	DWH	PP	SF	S
1	1	1	1	1	1
2	1	2	2	2	2
3	1	2	2	1	2
4	1	1	2	1	2
5	1	3	1	1	1
6	2	1	2	2	1
7	2	2	2	1	1
8	2	3	1	2	1
9	2	2	1	1	1
10	3	1	2	2	2
11	3	3	1	2	2
12	3	1	2	1	1
13	3	3	2	1	1
14	3	1	2	1	1
15	2	3	2	1	2

The procedure to find CCQ value is described here. To obtain the values of CCQ, firstly, we must obtain the equivalence classes induced by indisceribility relation of singleton attribute.

$X(\text{Demand}=1)=\{1,2,3,4,5\}$, $X(\text{Demand}=2)=\{6,7,8,9,15\}$, $X(\text{Demand}=3)=\{10,11,12,13,14\}$,

$$U/\text{Demand}=\{\{1,2,3,4,5\},\{6,7,8,9,15\},\{10,11,12,13,14\}\}.$$

$X(\text{DeleveryWH}=1)=\{1,4,6,10,12,14\}$, $X(\text{DeleveryWH}=2)=\{2,3,7,9\}$,

$X(\text{DeleveryWH}=3)=\{5,8,11,13,15\}$,

$$U/\text{DeleveryWH}=\{\{1,4,6,10,12,14\},\{2,3,7,9\},\{5,8,11,13,15\}\}.$$

$X(\text{ProductionPlan}=1)=\{1,5,8,9,11\}, X(\text{ProductionPlan}=2)=\{2,3,4,6,7,10,12,13,14,15\}$

$$U/\text{ProductionPlan}=\{\{1,5,8,9,11\},\{2,3,4,6,7,10,12,13,14,15\}\}.$$

$X(\text{Salesforcast}=1)=\{1,2,6,8,10,11\}, X(\text{Salesforcast}=2)=\{3,4,5,7,9,12,13,14,15\}$

$$U/\text{Salesforcast}=\{\{1,2,6,8,10,11\},\{3,4,5,7,9,12,13,14,15\}\}.$$

$X(\text{Supply}=1)=\{1,5,6,7,8,9,12,13,14\}; X(\text{Supply}=2)=\{2,3,4,10,11,15\}$

$$U/\text{Supply}=\{\{1,5,6,7,8,9,12,13,14\},\{2,3,4,10,11,15\}\}$$

Based on Definition 1, the error classification attribute Production Plan with respect to Demand is calculated as follow.

$$c(\text{Demand}=1,\text{ProductionPlan}=1)=1-\frac{|\{1,5\}|}{|\{1,2,3,4,5\}|}=1-\frac{2}{5}=\frac{3}{5},$$

$$c(Demand = 2, ProductionPlan = 1) = 1 - \frac{|\{8,9\}|}{|\{6,7,8,9,15\}|} = 1 - \frac{2}{5} = \frac{3}{5},$$

$$c(Demand = 3, ProductionPlan = 1) = 1 - \frac{|\{11\}|}{|\{10,11,12,13,14\}|} = 1 - \frac{1}{5} = \frac{4}{5},$$

$$c(Demand = 1, ProductionPlan = 2) = 1 - \frac{|\{2,3,4\}|}{|\{1,2,3,4,5\}|} = 1 - \frac{3}{5} = \frac{2}{5},$$

$$c(Demand = 2, ProductionPlan = 2) = 1 - \frac{|\{6,7,15\}|}{|\{6,7,8,9,15\}|} = 1 - \frac{3}{5} = \frac{2}{5},$$

$$c(Demand = 3, ProductionPlan = 2) = 1 - \frac{|\{10,12,13,14\}|}{|\{10,11,12,13,14\}|} = 1 - \frac{4}{5} = \frac{1}{5}.$$

By given $\beta = 0.2$, the quality of classification of the set of attribute Production Plan with respect to Demand as follows

$$\gamma = \frac{\sum_{U/a_i} | X_\beta(a_i = y_k)}{|U|}|$$

$$= \frac{|\{\phi\}| + |\{10,11,12,13,14\}|}{|15|}$$

$$= \frac{5}{15} = \frac{1}{3}.$$

Following the same procedure, the quality of classification on all attributes with respect each to the other are computed. These calculations are summarized in Table 2.
With CCQ technique, From Table 2, the highest quality of classification of attributes is Production Plan. Thus, attribute Production Plan is selected as a clustering attribute.

Table 2. The measure of classification quality of Table 1

Attribute (with respect to)	The quality of classification				mean
Demand	DWH	PP	SF	S	0
	0	0	0	0	
Delivery WH	D	PP	SF	S	0
	0	0	0	0	
Production Plan	D	DWH	SF	S	0.283
	0.333	0.4	0	0.4	
SalesForcast	D	DWH	PP	S	0.083
	0.333	0	0	0	
Supply	D	DWH	PP	SF	
	0.333	0	0.333	0	0.1665

For objects splitting, we use a divide-conquer method. For example, in Table 2 we can cluster (partition) the objects based on the decision attribute selected, i.e., Production Plan. Notices that, the partition of the set of animals induced by attribute Production Plan is

$$U / PP = \{\{1,5,8,9,11\}, \{2,3,4,6,7,10,12,13,14,15\}\}.$$

To this, we can split the objects using the hierarchical tree as follows.

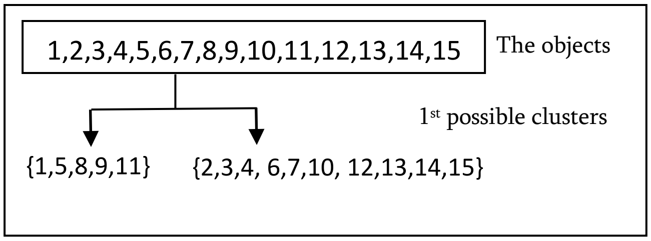

Figure 2. The objects splitting.

The technique is applied recursively to obtain further clusters. At subsequent iterations, the leaf node having more objects is selected for further splitting. The algorithm terminates when it reaches a pre-defined number of clusters. This is subjective and is pre-decided based either on user requirement or domain knowledge.

4. Experiment Results

We elaborate the proposed technique through the three UCI benchmark datasets taken from: Http:/kdd.ics. uci.edu. Balloon dataset contains 16 instances and 4 categorical attributes; Color, Size, Act and Age. Tic-Tac-Toe Endgame dataset The data contains 958 of instances and 9 categorical-attributes; top left square (TLS), top middle square (TMS), top right square (TRS), middle left square (MLS), middle middle square (MMS), middle right square(MRS), bottom left square (BLS), bottom middle square (BMS), bottom right square (BRS)and a class attribute. Hayes-Roth dataset contains 132 training instances, 28 test instances and 4 attributes; hobby, age, educational level and marital status. The algorithms of TR, MMR, and CCQ are implemented in MATLAB version 7.6.0.324 (R2008a). They are executed sequentially on a processor Intel Core 2 Duo CPUs. The total main memory is 1G and the operating system is Windows XP Professional SP3. The experiment results is summarized in Table 3.

The TR, MMR and CCQ use different techniques in selecting clustering attribute. TR uses the total average of mean roughness, MMR uses the minimum of mean roughness and CCQ uses the measure of classification quality of Variable Precision Rough Set to select a clustering attribute. Based on Table 3 the decision cannot be obtained using TR and MMR, because the value of TR and MMR of attributes in all datasets are same (for TR is

0 and for MMR is 1, respectively). But, the clustering attribute can be selected based on the highest values using CCQ.

Table 3. The experiment results

Technique	Data Set		
	Ballon	Tic tac toe	Hayes-Roth
TR	0	0	0
Attribute Selected	All	All	All
MMR	1	1	1
Attribute Selected	All	All	All
CCQ $(\beta = 0.4)$	0.8667	0.4541	0.3535
Attribute Selected	3 dan 4	5	3

The purity of clusters was used as a measure to test the quality of the clusters [9] The purity of a cluster and overall purity are defined as

$$Purity(i) = \frac{\begin{array}{c} Thenumberof\ datainboththeithcluster \\ anditscorrespondingclass \end{array}}{Thenumberof\ datainthedataset} \qquad (7)$$

$$OveralPurity = \frac{\sum_{i=1}^{\#\,of\ cluster} Putiy(i)}{\#of\ cluster}$$

We also use Rand Measure which is external validity to analyze the cluster The adjusted Rand index [12] is the corrected for chance version of the Rand index that computes how similar the clusters (returned by the clustering algorithm) are to the benchmark classifications. The Adjusted Rand Index is as follows

$$RI = \frac{\sum_{i=1}^{m}\sum_{j=1}^{K}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{m}\binom{n_{i.}}{2}\sum_{j=1}^{K}\binom{n_{.j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{m}\binom{n_{i.}}{2} + \sum_{j=1}^{K}\binom{n_{.j}}{2}\right] - \binom{n}{2}^{-1}\sum_{i=1}^{m}\binom{n_{i.}}{2}\sum_{j=1}^{K}\binom{n_{.j}}{2}} \qquad (8)$$

where n_{ij} represents the number of objects that are in predefined class i and cluster j, $n_{i\cdot}$ indicates the number of objects in a priori class i, $n_{.j}$ indicates the number of objects cluster j, and n is the total number of objects in the data set.

CR index takes its values from the interval [-1,1], in which the value 1 indicates perfect agreement between partitions, whereas values near 0 correspond to cluster agreement found by chance

Table 4. The cluster validity

	Data Set		
	Ballon	Tic tac toe	Hayes-Roth
Purity	0.83	0.69	0.63
Rank Index	66.3158	60.4557	54.0806

5. Conclusion

In this paper, we have proposed an alternative technique for categorical data clustering using Variable Precision Rough Set model. For selecting the clustering attribute, it is based on the classification quality of variable precision of attributes in the rough set theory. We present an example how our technique run. Further, we compare our technique on three benchmark datasets; Balloon and Tic-Tac-Toe Endgame and Hayes-Roth taken from UCI ML repository. The results show that our technique provides better performance in selecting the clustering attribute. Since TR and MMR are based on the traditional definition of rough set theory, thus our technique is different from TR and MMR.

References

1. Z. Huang, 1998 Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery 2 (3) 283–304.
2. R. Johnson, W. Wichern,. 2002. Applied Multivariate Statistical Analysis, Prentice Hall, New York.
3. In-Kyoo Park, Gyoo-Seok Choi, Rough set approach for clustering categorical data using information-theoretic dependency measure, Information Systems, Volume 48, March 2015, Pages 289-295, ISSN 0306-4379.
4. Min Li, Shaobo Deng, Lei Wang, Shengzhong Feng, Jianping Fan, Hierarchical clustering algorithm for categorical data using a probabilistic rough set model, Knowledge-Based Systems, Volume 65, July 2014, Pages 60-71, ISSN 0950-7051.
5. Pawlak, Z. 1982. Rough sets, International Journal of Computer and Information Science. 11, 341–356.
6. Pawlak, Z. 1991. Rough sets: A theoretical aspect of reasoning about data, Kluwer Academic Publisher.
7. Pawlak, Z. and Skowron, A. 2007. Rudiments of rough sets, Information Sciences 177 (1), 3–27.
8. Mazlack, L.J., He, A., Zhu, Y., Coppock, S. 2000. "A rough set approach in choosing partitioning attributes". Proceedings of the ISCA 13th, International Conference, CAINE-2000, 1–6.
9. Parmar, D., Wu, T.and Blackhurst, J. 2007. MMR: An algorithm for clustering categorical data using rough set theory, Data and Knowledge Engineering 63, 879–893.
10. Z. T. Gong, Z. H. Shi, H. Y. Yao. 2012. Variable Precision Rough Set Model For Incomplete Information Systems And Its Β-Reducts. Computing and Informatics, Vol. 31, 2012, 1385–1399
11. Ziarko, W. 1991.Variable precision rough set model, Journal of computer and system science 46, 39–59.
12. Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1), 193–218.