TELKOMNIKA, Vol. 13, No. 4, December 2015, pp. 1361~1367
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v13i4.2179

Alternative Technique Reducing Complexity of Maximum Attribute Relation

Iwan Tri Riyadi Yanto*[1], Imam Azhari[2]
Information System, Universitas Ahmad Dahlan (UAD)
3rd Campus, Jln. Prof. Soepomo, Janturan, Yogyakarta, Indonesia
*Corresponding author, email: yanto.itr@is.uad.ac.id[1], azhari@uad.ac.id[2]

Abstract

Clustering refers to the method grouping the large data into the smaller groups based on the similarity measure. Clustering techniques have been applied on numerical, categorical and mix data. One of the categorical data clustering technique based on the soft set theory is Maximum Attribute Relation (MAR). The MAR technique allows calculating all of pair multi soft set made. However, the computational complexity is still an issue of the technique. To overcome the drawback, the paper proposes the alternative algorithm to decrease the complexity so get the faster response time. In this paper, to get the similar results as MAR without calculating all pair of soft set is proved. The alternative algorithm is implemented in MATLAB Software, and then experimental is run on the 10 benchmark datasets. The results show that the alternative algorithm improves the computational complexity in term of response time up to 36.46%.

Keywords*: Soft set, Multi Soft set, Attribute Relation*

1. Introduction

Clustering is a fundamental problem that frequently arises in a broad variety of fields such as pattern recognition, image processing, machine learning and statistics [1, 2]. It can be defined as a process of partitioning a given data set of multiple attributes into groups. Clustering has been used in many areas such as gene data processing [3], transactional data processing [4], decision support [5], mobile ad-hoc networks (MANETs) [6], study anxiety [7] and radar signals processing [8]. Recently, many attentions have been put on categorical data clustering, where data objects are made up of non-numerical attributes [9, 10].

The main difference between categorical data and numerical data is the multi-valued attribute that belongs to the categorical data. These properties lead to difficulties in the similarities and dissimilarity measurement in the clustering process, since the normal distance measures cannot be applied directly to the categorical data. Therefore, the best similarity measurement of the categorical data is done by defining the common object for the attribute as well as the common values of the attribute, and the association between the two [9].

Currently, two measurement approaches based on the theory of rough set has been introduced in clustering attribute selection. The first approach is based on the roughness of the attribute, i.e. Total Roughness (TR) proposed by by Mazlack et al. [11] and Min– Min Roughness (MMR) proposed by Parmar et al. [12]. The second approach called Maximum Dependency of Attribute (MDA) proposed by Herawan et al. [13]. The approaches of finding a clustering attribute had successfully exploited the uncertainties in the multi-valued information system. But, there exists some unexpected iteration that leads to an increment in the processing time. the soft set theory proposed by Molodtsov is a new to manage uncertain data. Mamat et al. propose [14] MAR, an alternative technique to select a clustering attribute, One of the well known techniques based on soft set theory [15]. It is based on a concept of Maximum Attribute Relative where the comparison of attributes is made by taking into account the relative of the attribute at the category level after the multi-valued attribute is decomposed be a multi soft set. The proposed technique potentially discovers the attributes subsets with well coverage. However, the time complexity is still in issue since more categories on every attribute of a categorical data, the more multi soft set made. In this paper, the alternative algorithm to reduce the complexity of MAR is proposed. The difference with MAR is the proposed algorithm using Maximum Attribute Relative concept without calculating all of pair multi soft set made. The

Received December 23, 2014; Revised January 29, 2015; Accepted March 12, 2015

proposed algorithm potentially achieves a lower computational time complexity as compared to the original MAR.

2. Research Method

2.1. Information System

An information system (IS) refers to a collection of multiple pieces of equipment involved in the dissemination of information. Formally, as defined in [16] an information system can be represented as a 4-tuple (quadruple) $S=(U,A,V,f)$, where $U=\{u_1,u_2,u_3,\cdots u_{|U|}\}$ and $A=\{a_1,a_2,a_3,\cdots,a_{|A|}\}$ is a non- empty finite set of objects and attributes, respectively. $V=\bigcup_{a\in A} V_a, V_a$ is the domain (value set) of attribute a, $f:U\times A\rightarrow V$ is an information function such that $f(u,a)\in V_a$, for every $(u,a)\in U\times A$, called information (knowledge) function.

Definition 1 : let $S(U,A,V_{\{0,1\}},f)$ be an information system. If $V_a=\{0,1\}$, for every $\alpha\in A$, then $S(U,A,V_{\{0,1\}},f)$ is called a Bolean-valued information system.

2.2. Soft Set Theory

Let A is a set of parameter describing objects in $U, P(U)$ is the power set of U and $E\subset U$, soft set over U as defined in [15] is a pair (F,A), where F is a function given by

$$F: A\rightarrow P(U) \tag{1}$$

Obviously, a soft set (F,A) over U can be said as a parameterized family (subset) of the universe. For $\alpha\in A, F(\alpha)$ may be considered as the set of α-elements of the soft set $F(A)$ or the set α-approximate elements of the soft set $F(A)$.

Based on the definition of an information system and a soft set, as explained in [17], a soft set can be interpreted as a special type of information systems, termed a binary-valued information. The proposition and proof are given as follows

Propotition 1: Each soft set can be considered as a bolean-valued information system.

Proof : lets (F,E) be a soft set over universe $U, S=(U,A,V,f)$ be an information system. Obviously, the universe U in (F,E) can be considered as the universe U, the parameter set E can be considered as the attribute A. Then, the information system function, f is defined by

$$f=\begin{cases}1, & h\in F(e)\\ 0, & h\notin F(e)\end{cases} \tag{2}$$

That is, when $h_i\in F(e_j)$, where $h_i\in U$ and $e_j\in E$, then $f(h_i,e_j)=1$, otherwise $f(h_i,e_j)=0$. To this, we have $V(h_i,e_j)=\{0,1\}$. Therefore, a soft set (F,E) can be considered as Bolean-valued information system where $S=(U,A,V_{\{0,1\}},f)$ and a soft set (F,E) can be represented in the form of Bolean Table.

Definition 2: (see [18]) the class of all value sets of a soft set (F,E) is called value-class of the soft set and is denoted by $C_{(F,E)}$.

From proposition 1, the "standard" soft set deals with a Boolean-valued information system. For a categorical-valued information system $S=(U,A,V,f)$, where $V=\cup_{a\in A}V_a$, V_a is the domain (value set) of attribute a which has categorical (multi) values, a decomposition can be made from S into $|A|$ number of Boolean-valued information system $S=(U,A,V_{\{0,1\}},f)$. The decomposition of $S=(U,A,V,f)$ is based on decomposition of $A=\{a_1,a_2,\cdots,a_{|A|}\}$ into the disjoint-singleton attribute $\{a_1\},\{a_2\},\cdots,\{a_{|A|}\}$.

Definition 3: (see [14]) Lets $S=(U,A,V,f)$ be an information system such that for every $a\in A, V_a=f(U,A)$ is a finite non-empty set and for every $u\in U, |f(u,a)|=1$. For every a_i under i^{th}-attribute consideration, $a\in A$ and $v\in V_a$, the map a_v^i of U is defined as

$$a_v^i: U \to \{0,1\}, \tag{3}$$

such that

$$a_v^i(u) = \begin{cases} 1, & f(u,a) = v \\ 0, & otherwise \end{cases} \tag{4}$$

Definition 4: Lets $S = (U, A, V, f)$ be a categorical-valued information system and $S = \left(U, a_i, V_{a_i}, f\right), i = 1,2,\cdots,|A|$ Boolean-valued information system, we have

$$S = (U, A, V, f) = \begin{cases} S^1 = \left(U, a_1, V_{\{0,1\}}, f\right) \Leftrightarrow (F, a_1) \\ S^2 = \left(U, a_2, V_{\{0,1\}}, f\right) \Leftrightarrow (F, a_2) \\ \vdots \\ S^{|A|} = \left(U, a_{|A|}, V_{\{0,1\}}, f\right) \Leftrightarrow (F, a_{|A|}) \end{cases} = \left((F, a_1), (F, a_2), \cdots, \left(F, a_{|A|}\right)\right). \tag{5}$$

Then, $(F, E) = \left((F, a_1), (F, a_2), \cdots, \left(F, a_{|A|}\right)\right)$ can be defined as a multi soft set over universe U representing a categorical-valued information system $S = (U, A, V, f)$.

2.3. Maximum Attribute Relative (MAR)

The MAR technique approach has been proposed by Mamat et al [14]. There are three main steps to select the dominant attribute. The first step is determine the support of soft set respect to each parameters over the universe U. Consider to the pair (F, A), assign to multi-soft set over U, representing a categorical-valued information system $S = (U, A, V, f)$, where $(F, a_i), \cdots, \left(F, a_{|A|}\right) \subseteq (F, A)$ and $\left(F, a_{i_j}\right), \cdots, \left(F, a_{i_{|a_i|}}\right) \subseteq (F, a_i)$. Support of $\left(F, a_{i_j}\right)$ by $\left(F, a_{i_k}\right)$ denoted $sup_{\left(F, a_{i_k}\right)}\left(F, a_{i_j}\right)$ is defined as

$$sup_{\left(F, a_{i_k}\right)}\left(F, a_{i_j}\right) = \frac{|\left(F, a_{i_j}\right) \cap \left(F, a_{i_k}\right)|}{|\left(F, a_{i_k}\right)|}. \tag{6}$$

The next step is calculating the maximum and minimum support. The maximum support is defined as a summation of all support with value equal to 1. For each soft set $\left(F, a_{i_j}\right)$, the maximum support is denoted maxsup$\left(F, a_{i_j}\right)$, is defined as

$$\text{maxsup}\left(F, a_{i_j}\right) = \sum\left(sup_{\left(F, a_{i_k}\right)}\left(F, a_{i_j}\right) = 1\right). \tag{7}$$

Meanwhile, the minimum support is a summation of all support with value less than 1. For each soft set $\left(F, a_{i_j}\right)$, the maximum support is denoted minsup$\left(F, a_{i_j}\right)$, is defined as

$$\text{minsup}\left(F, a_{i_j}\right) = \sum\left(sup_{\left(F, a_{i_k}\right)}\left(F, a_{i_j}\right) \neq 1\right). \tag{8}$$

The MAR technique uses the highest and most frequently occured in the probability distribution.If $mode\left(\max\left(\text{maxsup}\left(F, a_{i_j}\right), \cdots, \text{maxsup}\left(F, a_{|m|_{|n|}}\right)\right) = 1\right)$, then $\left(F, a_{i_j}\right)$ is a clustering attribute. If $mode\left(\max\left(\text{maxsup}\left(F, a_{i_j}\right), \cdots, \text{maxsup}\left(F, a_{|m|_{|n|}}\right)\right) > 1\right)$ then $\max\left(\text{minsup}\left(F, a_{i_j}\right), \cdots, \text{minsup}\left(F, a_{|m|_{|n|}}\right)\right)$ is a clustering attribute, where max refers to the value that the highest in the probability distribution and mode refers to the value that is most frequently accored in the probability distribution.

```
Input  : Categorical data set
Output : Selected attribute
Begin
Builts the multi-soft set approximation
Calculate support, MaxSup and MinSup
for i=all categories
for j=all categories
Calculate intersection soft set i respect to soft set j
Calculate the support := intersection / soft set j
Calcukate as MaxSup or MinSup
end
end
Select attribute based on Maxsup and MinSup.
```

Figure 1. The pseudo code of MAR

2.4. Reducing Complexity

Throughout this section, a pair(F,A), refers to multi-soft sets over the universe U describing a categorical valued information system $S=(U,A,V,f)$. Consider to the support in [14], the value of $sup_{\left(F,a_{i_k}\right)}\left(F,a_{j_l}\right) \in [0,1]$ is clear. However, the value of support is always 0 or 1 in the certain case. The justification is given in proposition 2.

Proposition 2: Lets $S=(U,A,V,f)$ be an information system, represented as a pair (F,A) as multi soft set over U, where $(F,a_i),,\cdots,\left(F,a_{|A|}\right) \subseteq (F,A)$ and $\left(F,a_{i_1}\right),\cdots,\left(F,a_{i_{|a_i|}}\right) \subseteq (F,a_i)$. If

$$sup_{\left(F,a_{i_k}\right)}\left(F,a_{j_l}\right)=\frac{|\left(F,a_{j_l}\right)\cap\left(F,a_{i_k}\right)|}{|\left(F,a_{i_k}\right)|}, for\ i=j. \quad (9)$$

then

1. $sup_{\left(F,a_{i_k}\right)}\left(F,a_{j_l}\right)=0, for\ k\neq l$
2. $sup_{\left(F,a_{i_k}\right)}\left(F,a_{j_l}\right)=1$, for $k=l$

Proof.

1. Let (F,A) be a multi soft set over the universe U, where $(F,a_i),,\cdots,\left(F,a_{|A|}\right) \subseteq (F,A)$ and $\left(F,a_{i_1}\right),\cdots,\left(F,a_{i_{|a_i|}}\right) \subseteq (F,a_i)$. Based on the definition 4, the multi soft set is constructed by decomposing the multi valued of attribute in the information system $S=(U,A,V,f)$ be multi parameters. In the other words, each attribute can be reconstructed into some pair of soft set i.e: $\left(F,a_{i_1}\right)\cup\cdots\cup\left(F,a_{i_{|a_i|}}\right)=U$ and $\left(F,a_{i_l}\right)\cap\left(F,a_{i_k}\right)=\emptyset$, for $k\neq l$. Thus, for $i=j, sup_{\left(F,a_{i_k}\right)}\left(F,a_{j_l}\right)=0$,when $k\neq l$.
2. Clear. The soft set intersects with its self.

Based on the proposition 2, the support value of each attributes is exactly always 1 if the soft set intersects with itself and 0 if the soft set intersects respect to other soft set in the same attribute. For this reason, not all support of pair sets in the multi soft set (F,A) is calculated, but we only calculate the support of pairs soft set respect to all soft set in the other attribute. Therefore, the complexity of algorithm will be reduced. Suppose that in an information system, there are n objects, m attributes and l is the maximum distinct values of each attribute. Computational cost to determine the elementary set of all attributes is $nm+1$. The proposed technique needs $ml(ml-1)$ times to determine the support for each category. Thus, the computational complexity for the MAR technique is $O\left(ml(ml-1)+nm+1\right)=O\left((ml)^2+(n-l)m+1\right)$. however, the modification algorithm needs $ml(m-1)l$. Thus, the computational complexity for the proposed technique is $O\left(ml(m-1)l+nm+1\right)=O\left((ml)^2+(n-l^2)m+\right.$

1). The pseudo code of MAR and alternative algorithm are given in figure 1 and figure 2, respectively.

```
Input  : Categorical data set
Output : Selected attribute
Begin
Builts the multi-soft set approximation
Calculate support, MaxSup and MinSup
for i=all attribute
for j=all attribute
if i isnot equal to j
Calculate intersection all soft set in attribute i respect to all soft set in attribute j
Calculate the support := intersection /  each soft set in the attribute j
Calculate the minsup and maxsup
end
end
Select attribute based on Maxsup and MinSup.
```

Figure 2. The pseudo code of proposed algorithm

3. Experimental Results and discussion

In order to compare the proposed algorithm and MAR approach, the both algorithms are implemented in MATLAB win 32 bit version 7.10.0 (R2010a). They are executed sequentially on a processors Intel Atom Quad core @1.50 GHz (4 CPUs). The total main memory is 2G and the operating system is Windows 7 Professional 32-bit. We elaborate through the UCI benchmark datasets [19] as in table 1:

Table 1. The UCI benchmark datasets

Datasets	Number of Instances	Number of Attributes	Size of Data
AcuteImflammations	120	6	720
balloon	16	5	80
cilinder_band	520	14	7,280
lenses	24	5	120
lungCancer	32	57	1,824
mushroom	8,124	22	178,728
solar_flare	1,066	13	13,858
soybean	47	35	1,645
soybean_large	265	36	9,540
suplier	27	7	189

All the selected data sets are different from one another in terms of size, either horizontally or vertically aimed to analyze the performance of the proposed technique when involving a high number of records as well as the high number of attributes. Some datasets have been modified by removing instances having incomplete data and removing an attribute only having one categorical value.

Table 2. Computation time of MAR and proposed algorithm

Datasets	TR	MMR	MDA	MAR	The proposed algorithm	Improvement
AcuteImflammations	0.3038	0.3080	0.2896	0.1880	0.1560	17.02%
balloon	0.0171	0.0194	0.0194	0.0160	0.0150	6.25%
cilinder_band	0.2231	0.2409	0.1175	0.1090	0.0620	43.12%
lenses	0.1085	0.1240	0.0310	0.0310	0.0160	48.39%
lungCancer	0.1395	0.1292	0.0775	0.0620	0.0470	24.19%
mushroom	1.0128	1.1517	1.0060	0.0470	0.0200	57.45%
solar_flare	0.4694	0.4687	0.3957	0.0780	0.0470	39.74%
soybean	0.0858	0.0633	0.0349	0.0310	0.0160	48.39%
soybean_large	0.2306	0.2490	0.1214	0.1090	0.0780	28.44%
suplier	0.0395	0.0550	0.0352	0.0310	0.0150	51.61%
Average						36.46%

The computation results comparing the MAR and proposed algorithm in term of execution time are shown in table 2. A decreasing relative velocity of the proposed algorithm respect to MAR is calculated by following formula:

$$Impr(\%) = \frac{(MAR - the\ proposed\ algorithm)}{MAR} \times 100\% \quad (10)$$

In summary, based on experiments on UCI datasets, the proposed algorithm achieves lower computation time than the previous algorithm. An increase average time of proposed algorithm reach 36.46 %.

Balloon dataset containing 16 objects, 5 attributes and suplier containing 27 objects, 7 attributes are the fastest execution times, standing at 0.015 second at proposed algorithm, 0.016 second and 0,031 second at MAR, improving up to 6.25 % and 51.61 %, respectively. A part from that, the longest execution time is at acute inflammations data set containing 120 objects and 6 attributes, the proposed algorithm needs 0.156 second, improving 17.02 % from 0.188 second needed MAR. The highest improvement achieves 57.45 % on the mushroom datasets, comprising 8124 objects and 22 attributes. On other hand, the lower improvement is on balloon datasets. The range of the highest and the lowest improvement is 51.20% while the average improvement reaches 36.46% of 10 datasets.

4. Conclusion

A number algorithm used to selecting attribute on categorical data clustering problem have been proposed, one of them is MAR. However, computational complexity of this algorithm is still an issue. This paper proposes an alternative algorithm modified MAR based on multi soft set theory selecting attribute to clustering multi-value information systems. The modification allows not all multi soft set composed attributes calculated. The experiment results illustrate the proposed algorithm achieve lower execution time. The main contribution of this work is in terms of reducing execution time where it is slightly improved as compared to MAR. In the next stage, the ability of the technique to classify categorical data will be examined in terms of clustering efficiency and accuracy of clustering. It also needs to be developed for grouping different kinds of data types.

References

[1] Z Huang. “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values”. *Data Min. Knowl. Discov.* 1998; 2(3): 283–304.

[2] DW Kim, KH Lee and D Lee. “Fuzzy clustering of categorical data using fuzzy centroids”. *Pattern Recognit. Lett.* 2004; 25(11): 1263–1271.

[3] D Jiang, C Tang and A Zhang. “Cluster analysis for gene expression data: a survey”. *IEEE Trans. Knowl. Data Eng.* 2004; 16(11): 1370–1386.

[4] F Giannotti, C Gozzi and G Manco. “Clustering Transactional Data”. in *Principles of Data Mining and Knowledge Discovery SE - 15*, T Elomaa, H Mannila and H Toivonen, Eds. Springer Berlin Heidelberg. 2002; 2431: 175–187.

[5] RG Mathieu and JE Gibson. “A methodology for large-scale R&D planning based on cluster analysis”. *IEEE Trans. Eng. Manag.* 1993; 40(3): 283–292.
[6] A Afsharfarnia and A Karimi. “A New Clustering Algorithm Using Links’ Weight to Decrease Consumed Energy in MANETs”. *TELKOMNIKA (Telecommunication Comput. Electron. Control.* 2014; 12(2): 411.
[7] ITR Yanto, P Vitasari, T Herawan and MM Deris. “Applying variable precision rough set model for clustering student suffering study’s anxiety”. *Expert Syst. Appl.* 2012; 39(1): 452–459.
[8] S Haimov, M Michalev, A Savchenko and OI Yordanov. “Classification of radar signatures by autoregressive model fitting and cluster analysis”. *IEEE Trans. Geosci. Remote Sens.* 1989; 27(5): 606–610.
[9] S Deng, Z He and X Xu. “G-ANMI: A mutual information based genetic clustering algorithm for categorical data”. *Knowledge-Based Syst.* 2010; 23(2): 144–149.
[10] K Chen and L Liu. “‘Best K’: critical clustering structures in categorical datasets”. *Knowl. Inf. Syst.* 2008; 20(1): 1–33.
[11] AHYZ Lawrence J Mazlack. “A Rough Set Approach in Choosing Partitioning Attributes”.
[12] D Parmar, T Wu, and J Blackhurst. “MMR: An algorithm for clustering categorical data using Rough Set Theory”. *Data Knowl. Eng.* 2007; 63(3): 879–893.
[13] T Herawan, MM Deris and JH Abawajy. “A rough set approach for selecting clustering attribute”. *Knowledge-Based Syst.* 2010; 23(3): 220–231.
[14] R Mamat, T Herawan and MM Deris. “MAR: Maximum Attribute Relative of soft set for clustering attribute selection”. *Knowledge-Based Syst.* 2013; 52: 11–20,.
[15] D Molodtsov. “Soft set theory—First results”. *Comput. Math. with Appl.* 1999; 37(4–5): 19–31.
[16] MK Ng. “A fuzzy k-modes algorithm for clustering categorical data”. *IEEE Trans. Fuzzy Syst.* 1999; 7(4): 446–452,.
[17] T Herawan. “A Direct Proof of Every Rough Set is a Soft Set”. in *Asia International Conference on Modelling & Simulation*. 2009; 0: 119–124.
[18] MI Ali, F Feng, X Liu, WK Min and M Shabir. “On some new operations in soft set theory”. *Comput. Math. with Appl.* 2009; 57(9): 1547–1553.
[19] UCI Machine Learning Repository:, “https://archive.ics.uci.edu/ml/datasets.html”.