Cluster Validity Analysis on Soft Set Based Clustering

Rabiei Mamat[1], Mustafa Mat Deris[2], Ahmad Shukri Mohd Noor[1] and Sumazly Sulaiman[1]
[1]*School of Informatics & Applied Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu Malaysia.*
[2]*Faculty of Computer Science & Information Technology, Universiti Tun Hussein Onn Malaysia,*
86400 Parit Raja, Batu Pahat, Johor, Malaysia.
rab@umt.edu.my

***Abstract*—The issue of data uncertainties are very important in categorical data clustering since the boundary between created clusters are very arguable. Therefore the algorithm called Maximum Attribute Relative (MAR) that is based on the attribute relative of soft-set theory was proposed previously. MAR exploiting the data uncertainties in multi-value information system by introducing a series of clustering attribute. The clusters will be form by using this selected clustering attributes. However, clustering algorithm define clusters that are not-known a priori. Hence, the final clusters of data requires some validation techniques. In this paper, the validity of the clusters produced by MAR was evaluated. The two datasets obtained from UCI-ML repository and an examination results obtained from Malaysian Ministry of Education. The results shows that the clusters produced by MAR has objects similarity up to 99%.**

***Index Terms*— Attribute Relative; Categorical Data; Data Clustering; Soft-Set.**

I. INTRODUCTION

Data clustering is an activity of grouping data into similar group based on some properties. The results is groups of data (clusters) that are similar to each other within group with respects to the properties and the groups themselves stand apart from one another. In other word, the objective is to divide the data into homogenous and distinct groups. But, when involving the categorical data, the categorical variables are very hard to measure and leads to the difficulties in determining the objects similarity resulting to the uncertainties in data. According to Molodtsov, the reasons of this difficulties is due to the inadequacy of the parameterized tools used. Molodtsov then initiated a new mathematical tool called a soft-set theory [1] which claims to have adequate parameterized tools for dealing with uncertainties. The soft-set used parameterization sets as it main solution for problem solving which makes it very convenient and easier to apply.

Based on the theory of soft-set, a new algorithm for clustering categorical data was proposed. The algorithm which is called Maximum Attribute Relative (MAR) built to exploits the uncertainties in the categorical data by introducing the series of clustering attribute. The algorithm already shows a better processing time when applied using some datasets from UCI-ML [2].

However, a good clustering algorithm does not depend only on better processing time, but most important is the validity of clustering results. The validation of the clustering result or also known as cluster validity analysis is the assessment of a clustering procedure's output. But, the result of different clustering algorithm on the same dataset are varies since they are bounded to the input parameters and the behavior of the algorithms itself. Therefore, the precise technique of cluster validity measures must be determined which later will reflects the definition of a good clustering scheme.

In this paper, we evaluated the MAR technique cluster validity. In Section II, the definition of soft-set theory and the important definition of multi-soft set is explained and elaborated. Also in this section, the relationship between the soft-set theory and the information system is discussed together with the definition of binary information system. In Section III, the definition of maximum attribute relative of soft-set theory is explained. It is the follow by the discussion about the cluster validation in Section IV where three methods external validation is explained. In Section V, the cluster validation result is explained and discussed. Finally, the conclusion is given in Section VI.

II. SOFT-SET THEORY AND INFORMATION SYSTEM

The earlier idea of soft-set can be traced to the work of Pawlak in [3], where the Pawlak's concept of soft-set theory is a unified view of classical set, rough set and fuzzy set. However, today's soft-set theory is a result of Molodtsov's paper entitled "Soft set: a first result" [1] where the notion of soft-set theory has been defined. Molodtsov's notion of soft-set theory is a general method for dealing with uncertain that is free from the inadequacy of the parameterization tools. Molodtsov also presented some applications of the soft-set theory in several directions such as in game theory and operation research. According to Molodtsov in [1], if given an initial universe called U which contains a collection of objects that was described by parameter E, then exist the power set of U which is denoted by $P(U)$.

Definition 1: [1] If $A \subseteq E$, a pair (F, A) is called a soft-set over U if and only if F is mapping of A into the set of all subsets of the universe U. Mathematically, the definition is as in the equation (1).

$$F: A \to P(U) \text{or } A \xrightarrow{f} P(U) \tag{1}$$

By definition 1, it is clear that a soft-set over the universe U is referred to any subset of U parameterized by E. Thus, for a given $\alpha \subseteq E$, $F(\alpha)$ is considered as an approximation of soft-set (F, E) parameterized by α. In other words, the soft-set is a parameterized family of subsets of the set U.

Example 1: Let universe $U = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}\}$ is a set of candidates for hire described by a set of soft-skills $E = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7\}$ where each soft-skills $e_i \in E$ respectively stand for communicative, critical thinking, teamwork, information management, entrepreneurship, leadership and morale. Suppose that each candidates has the skills as follows: $c_1 = \{e_1\}$, $c_2 = \{e_1, e_4, e_5\}$, $c_3 = \{e_2, e_4\}$, $c_4 = \{e_1, e_4\}$, $c_5 = \{e_1, e_4, e_5\}$, $c_6 = \{e_3, e_5, e_6, e_7\}$, $c_7 = \{e_5\}$, $c_8 = \{e_2, e_4, e_5\}$, $c_9 = \{e_2, e_3, e_5, e_6, e_7\}$ and $c_{10} = \{e_3, e_5, e_6, e_7\}$. Therefore defining a soft-set (F, E) as a subset of the universe U parameterized by $e_i \in$ E will give us a collection of approximate description of an object. In this case, a soft-set that describes "the capabilities of the candidate to be hired" is defined by communication, critical-thinking and so-on. If the mapping F is "candidate (*)" where (*) is to be filled by a parameter $e_i \in E$, then $F(e_1)$ means candidates (communication) whose functional value is the set $\{c_1, c_2, c_4, c_5\}$. Obviously, an approximation of soft-set has two parts: a predicate and an approximation value-set. In above case, the predicate is "communication" and the approximate value-set is $\{c_1, c_2, c_4, c_5\}$. Thus the soft-set (F, E) can be viewed as a collection of approximation as in the following figure.

$$(F,E) = \left\{ \begin{matrix} F(e_1) = \{c_1, c_2, c_4, c_5\} \\ F(e_2) = \{c_3, c_8, c_9\} \\ F(e_3) = \{c_6, c_9, c_{10}\} \\ F(e_4) = \{c_2, c_3, c_4, c_5, c_8\} \\ F(e_5) = \{c_2, c_5, c_6, c_7, c_8, c_9, c_{10}\} \\ F(e_6) = \{c_6, c_9, c_{10}\} \\ F(e_7) = \{c_6, c_9, c_{10}\} \end{matrix} \right\}$$

Figure 1: Approximation of soft-set (F, E)

The relationship between soft-set and information system has been an attention among the researcher around the globe. Pei and Miao [4] had insisted that soft-set and information system had a compact connection between them where soft-set is classified as a special class of information system.

Definition 2: An information system S is defined as a quadruple (U, A, V, F) where $U = \{x_1, \cdots, x_n\}$ is a non-empty finite set of interested objects, $A = \{a_1, \cdots, a_m\}$ is a non-empty finite set of attributes, $V = \bigcup_{i=1}^{m} V_i$ where V_i is the value set of the attribute a_i and $F = \{f_1, \cdots, f_m\}$ is an information function where $f_i: U \times a_i \to V_i$ such that $f(x, a) \in V_a$ for every $(x, a) \in U \times A$.

The information system is called classical information system when every V_i only contains finite elements either the elements is a number or not, for every $i \leq m$. However, if $V_i = [0,1]$ for every $i \leq m$, then the corresponding information systems are called fuzzy information systems. Furthermore, if $f_i: U \to P(V_i)$ is a mapping from U to the power set of V_i for all $i \leq m$ then the corresponding information system is called set-valued information system. The precise concept of an information system can be found in [5,6,7,8,9].

Property 1: Each soft-set can be considered as a Boolean-valued information system.

Proof. Let(F, E) be a soft-set over the universe O and $S = (U, A, V, f)$ be an information system. Let consider the universe O is a universe U in information system S and the parameter set E can be considered as the attributes A in S. Then, the information function f is defined by

$$f(u_i, e_j) = \begin{cases} 0, & u_i \in F(e_j) \\ 1, & u_i \in F(e_j) \end{cases}. \quad (2)$$

By equation (2), the $f(u_i, e_j)$ is set to 1,$\forall u_i \in U$, when $u_i \in F(e_j)$ and $e_j \in E$. Otherwise, $f(u_i, e_j)$ is set to 0, $\forall u_i \in U$ and $e_j \in E$. Clearly it shown that $V(u_i, e_j)$ is limited to $\{0,1\}$. Therefore, a soft-set (F, E) can be considered as a Boolean-valued information system where$S = (U, A, V_{\{0,1\}}, f)$.

Example 2: Let consider the approximation of soft-set (F, E) as in Fig 1. By taking object $c_1 \in U$ and parameter $e_1 \in E$ as an input to the function *f*, then the output is '1' since $c_1 \in F(e_1)$ but if taking $c_1 \in U$ and $e_2 \in E$, the output is '0' since $c_1 \notin F(e_2)$. As a result, a soft-set (F, E) can be represented in the form of information system $S = (U, A, V, f)$ as shown in Table 1. It can be seen in the table, as check against the Fig 1, '1' denoted the presence of the described parameters, while '0' means the parameter is not part of the capabilities of the candidate to be hired.

Table1
Information system built from soft-set $(\boldsymbol{F}, \boldsymbol{E})$ approximation

U	e_1	e_2	e_3	e_4	e_5	e_6	e_7
c_1	1	0	0	0	0	0	0
c_2	1	0	0	1	1	0	0
c_3	0	1	0	1	0	0	0
c_4	1	0	0	1	0	0	0
c_5	1	0	0	1	1	0	0
c_6	0	0	1	0	1	1	1
c_7	0	0	0	0	1	0	0
c_8	0	1	0	1	1	0	0
c_9	0	1	1	0	1	1	1
c_{10}	0	0	1	0	1	1	1

Unfortunately, the structure of standard soft-set is very simple. As can be seen in the Table 1, the mapping only classifies the objects into two classes either '1' or '0'. But, in the real world, attributes in an information system may have more than two properties or called multi-valued information system. Herawan et.al introduced multi-soft set[10] to overcome the issue. The composition of multi-soft set is make by decomposition of $A = \{a_1, \cdots, a_{|A|}\}$ from multi-valued information system $S = (U, A, V, f)$ into disjoint singleton attribute $\{a_1\}, \cdots, \{a_{|A|}\}$. For every a_i under i^{th}-attribute consideration, where $a_i \in A$ and $v \in V_a$, $a_v^i: U \to \{0,1\}$ such that $a_v^i(u) = 1$ if $f(u, a) = v$, otherwise $a_v^i(u) = 0$. The summary of the soft-set decomposition is shown as the following:

$$S=(U,A,V,f)=\begin{cases}S^1=(U,a_1,v_1,f)\Leftrightarrow(F,a_1)\\ \vdots\\ S^{|A|}=(U,a_{|A|},v_{|A|},f)\Leftrightarrow(F,a_{|A|})\end{cases}$$

where, $(F,a_1)=\begin{cases}S_0^1=(U,a_{1_0},v_{\{0,1\}},f)\Leftrightarrow(F,a_{1_0})\\ \vdots\\ S_{|a_1|}^1=(U,a_{|a_1|},v_{\{0,1\}},f)\Leftrightarrow(F,a_{|a_1|})\end{cases}$

and $(F,a_{|A|})=\begin{cases}S_0^{|A|}=(U,a_{|A|_0},v_{\{0,1\}},f)\Leftrightarrow(F,a_{|A|_0})\\ \vdots\\ S_{a_{|A|}}^{|A|}=\left(U,a_{|A|_{|A|}},v_{\{0,1\}},f\right)\Leftrightarrow\left(F,a_{|A|_{|A|}}\right)\end{cases}$.

Thus, $(F,E)=\left((F,a_1),\cdots,\left(F,a_{|A|}\right)\right)$.

The result of multi-soft set decomposition $(F,E)=\left((F,a_1),\cdots,\left(F,a_{|A|}\right)\right)$ is defined as a multi-soft set over universe Urepresenting a multi-valued information system$=(U,A,V,f)$.

III. Attribute Relative of Soft Set Theory

In this section, some definition that is related to the attribute relative of soft-set is given. Throughout this section a pair(F,A) is refer to multi-soft sets over the universe U that representing a categorical-valued information system $S=(U,A,V,f)$.

Definition 3: Let$\left((F,a_0),\cdots,\left(F,a_{|A|}\right)\right)\subseteq(F,A)$ and $\left(\left(F,a_{0_0}\right),\cdots,\left(F,a_{\lceil a_0\rceil}\right)\right)\subseteq(F,a_0)$, And let $\left(\left(F,a_{[A]_0}\right),\cdots,\left(F,a_{|A|_{a_{|A|}}}\right)\right)\subseteq\left(F,a_{|A|}\right)$. A soft-set$\left(F,a_{j_k}\right)$ is said to be relative to$\left(F,a_{p_q}\right)$ and vice-versa if $\left(F,a_{j_k}\right)\cap\left(F,a_{p_q}\right)\neq\emptyset$.

Definition 4: If a soft-set $\left(F,a_{p_q}\right)$ is relative to$\left(F,a_{j_k}\right)$, then the relative support value of$\left(F,a_{p_q}\right)$ by $\left(F,a_{j_k}\right)$which is denoted by $RSup_{\left(F,a_{j_k}\right)}\left(F,a_{p_q}\right)$ is defined as Equation (3).

$$RSup_{\left(F,a_{j_k}\right)}\left(F,a_{p_q}\right)=\frac{\left|\left(F,a_{p_q}\right)\cap\left(F,a_{j_k}\right)\right|}{\left|\left(F,a_{j_k}\right)\right|}. \quad (3)$$

Result of relative support value calculation can be categorized into either full relative, partly relative or zero (no) relative as describe by equations (4).

$$RSup\ Value=\begin{cases}1 & ,full\ relative\\ >0\ and\ <1 & ,partly\ relative\\ 0 & ,zero\ relative\end{cases} \quad (4)$$

Definition 5: Total relative support is a summation of all full relative support for each soft-set $(F,a_i)\subseteq(F,a_i)\subseteq(F,A)$. Hence, the total relative support of soft-set $\left(F,a_{i_m}\right)$ which is denoted by $TRS_{\left(F,a_{i_m}\right)}$ is computed by the following eqution (5):

$$TRS_{\left(F,a_{i_m}\right)}=\sum_{i=0,j=0}^{i=|A|,j=|a_i|}\left(Full\ Relative_{\left(F,a_{i_j}\right)}\left(F,a_{i_m}\right)\right) \quad (5)$$

Definition 6: Total attribute relative is a summation of all total relative support for each soft set$(F,a_i)\subseteq(F,A)$. Equation (6) show how the total attribute relative for (F,a_0) which is denoted by$TAR_{(F,a_0)}$ is calculated.

$$TAR_{(F,a_0)}=\sum_{i=0,j=0}^{i=|A|,j=|a_i|}TRS_{\left(F,a_{i_j}\right)} \quad (6)$$

Definition 7: Max is referred to the value that is the highest in the probability distribution

Definition 8: Maximum attribute relative is the maximum value of TAR in the probability distribution which is denoted by MAR as shown in Equation (7).

$$MAR=Max\left(TAR_{(F,a_0)},\cdots,TAR_{(F,a_{|A|})}\right). \quad (7)$$

Definition 9: Mode refers to the value that is most frequently occurred in the probability distribution.

Property 2: If $MAR=Max\left(TAR_{(F,a_0)},\cdots,TAR_{(F,a_{|A|})}\right)=1$, then (F,a_i) is a partition attribute.

Proof: If (F,a_i) is the maximum of attribute relative then it is obvious that (F,a_i)have more full relative as compared to others. Thus, from definition 3, 4, 5, 6, 7, 8 and 9, (F,a_i) is selected as a clustering attribute.

Corollary 1: If $Mode\left(Max\left(MAR\left((F,a_i),,\left(F,a_{|A|}\right)\right)\right)\right)>1$, then the clustering attribute is $Max\left(TRS_{MAR_0},,TRS_{MAR_{|A|}}\right)$.

Proof: Let (F,a_i) and$\left(F,a_j\right)$, be two (2) soft set over the universe $U$and let (F,a_i) and $\left(F,a_j\right)$ is a member ofMAR, and $TAR(F,a_i)=TAR\left(F,a_j\right)$ are maximum. Both attributes cannot be used as a clustering attribute unless it is proven that both attribute have the same full relative support attribute which can only be proved by the TRS value at the categorical level. Hence, if $TRS_{(F,a_i)}$of MAR is maximum then it is clear that $TRS_{(F,a_i)}$ is most relative to all other categorical soft-set and is selected as the clustering attribute.

IV. Cluster Validation

According to Sripada and Rao [11], there are two types of cluster validations measures; internal validation and external validation. The internal validations measures use the information that is intrinsic to the data-set to measure the quality of the obtained clusters. Meanwhile, the external validations examine the output of the clustering process against an existing set of class label of the data-set in determining the degree of occurrences according to that class label. For most applications, the external cluster validations measures are much more appropriate. One of the popular external cluster validations measures is Entropy, which is refer to the Shannon entropy [12]. Entropy was developed to measure the uncertainty associated with a random variable. For the cluster analysis, entropy measures the quality of the cluster with respect to the given class labels or in other word, entropy measure the distribution of various clusters within

each class. Entropy method has been used in measuring the validity of HICAP [13], comparing K-Means and fuzzy C-Means and measuring hierarchical clustering document algorithm for the document dataset in [14]. Then [15] uses this measure to evaluate the performance of their alternate least-square NMF algorithm.

For a cluster validation, the entropy is a summation of the class distribution of the objects in each cluster. Let's consider i is the number of class and j is the number of cluster. The distribution of the objects in each cluster is the probability that a member of cluster j belongs to class$i$ denoted by p_{ij}. Then, the entropy of cluster j denoted by E_j is calculated using the standard entropy formula as shown in Equation (8).

$$E_j = -\sum_i p_{ij}\, log_2(p_{ij}). \tag{8}$$

So, the total entropy for a set of clusters is computed as the weighted sum of the entropies of each cluster denoted by E as the following Equation (9),

$$E = \sum_{j=1}^{m} \frac{n_j}{n} \times E_j \tag{9}$$

where $n_j$is the size of cluster j, m is the number of clusters and n is the total number of data points. The entropy value near to zero (0) is interpreted as a better clustering, otherwise when the entropy values near to one (1), the quality of clustering is in doubt. By the other means, the lower entropy shows that the method used in the clustering process have successful reduce or managed the uncertainties among data. Otherwise, the uncertainties are not well organized by the method.

Meanwhile, Hubert and Arabie [16] introduced Adjusted Rand Index (ARI), which aiming to establish an overall comparison between the computed cluster and their equivalent class label. It is based on the Rand Index [17] which compares each pair assignment in the class labels and in the computed cluster. In other words, ARI measured an agreement between the computed clusters and the class labels. ARI have been used in [18] for clustering gene expression data and also applied in [19] as a performance measure in supervised classifications. Let S be a set of d data points where $S = \{o_1, o_2, \cdots, o_{n-1}, o_n,\}$. Given two (2) clustering of S namely $P = \{p_1, \cdots, p_m\}$ is a computed cluster with M clusters and suppose $C = \{c_1, \cdots, c_n\}$ is an ideal cluster with N clusters such that $\cup_{i=1}^{M} P_i = S = \cup_{j=1}^{N} C_j$ and $p_i \cap p_{i'} = 0 = c_j \cap c_{j'}$ for $1 \le i \neq\neq i' \le M$ and $1 \le j \neq j' \le N$. The information about cluster overlaps between P and C can be summarized in the form of a $R \times C$contingency table $K = [n_{ij}]_{j=1,\cdots,C}^{i=i,\cdots,R}$ as the following Table 2.

Table 2
Contingency table for partition overlapping

P/C	C_1	$\cdots$	C_N	Total
P_1	n_{11}	$\cdots$	n_{1N}	a_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
P_M	n_{M1}	$\cdots$	n_{MN}	a_M
Total	b_1	$\cdots$	b_N	

where n_{ij} is the number of $P_i \cap P_j$, a_i and b_j respectively is summation of row i and column j in the table. Based on the contingency Table 3.1, let $w = \binom{d}{2}, x = \sum_{i=0,j=0}^{M,N} \binom{n_{ij}}{2}, y = \sum_{i=0}^{M} \binom{a_i}{2}$ and $z = \sum_{j=0}^{N} \binom{b_j}{2}$ then ARI is computed using the Equation (10).

$$ARI = \frac{x - (a.b)/w}{\frac{(a+b)}{2} - \frac{(a.b)}{w}} \tag{10}$$

Another external cluster validation measure in concern is F-Measure [20] which uses the combination of Precision and Recall, two (2) concepts used in the information retrieval. It is usually a preferred accuracy standard performance measured in information retrieval especially when relevant items are rare. It has been used in [21] to evaluate unsupervised clustering with non-determined number of clusters. The Recall and Precision for each cluster of each given class is calculated using the Equation (11) and (12) respectively,

$$Recall(i,j) = \frac{n_{ij}}{n_i}, \tag{11}$$

$$Precision(i,j) = \frac{n_{ij}}{n_j} \tag{12}$$

where n_{ij} is the number of objects in cluster i that is in the class j, n_i is the number of objects in the cluster i and n_j is the number of objects in class j. Therefore, the F-Measure of cluster i and class j is then computed as using Equation (13).

$$F - Measure(i,j) = \frac{2 \times Precision(i,j) \times Recall(i,j)}{Precision(i,j) + Recall(i,j)} \tag{13}$$

But, the cluster validity is measure by using the weighted average of F-Measure for each class which computed as in Equation (14),

$$FM_{(WA)} = \sum_i \left(\frac{n_i}{n}\right) Max\big(F - Measure(i,j)\big), \tag{14}$$

where Max is taken over all clusters at all levels and n is the number of class. $FM_{(WA)}$ values are in the interval $[0,1]$ whereas larger values indicates higher clustering quality.

V. Cluster Validations Analysis and Result

Cluster validation analysis is carried-out using two standard datasets from UCI-MLR i.e. Soybean (Small) and Zoo, and a dataset from Malaysian Ministry of Education. As for a validity comparison, the results produce by MAR technique was compared to the result produce by MDA technique [22], a technique that uses the same approach as MAR but based on the theory of rough set.

In summary, the soybean (small) dataset contains forty-seven (47) objects of soybean on diseases. The dataset is in the completed state without any missing values. Each object is classified into one (1) of the four (4) diseases either *Diaporthe Stem Canker* (D1), *Charcoal Rot* (D2), *Rhizoctonia Root Rot* (D3) or *Phyrophthora Rot* (D4). The dataset is comprised of ten (10) objects of D1, ten (10) objects of D2, 10 (ten) objects of D3 and seventeen (17) objects of

D4. The Entropy, ARI and F-Measure results by using soybean (small) for both techniques is given in Table 3, while the graph that depicting the validity of MAR as compare to MDA is shown in Figure 2.

Table 3
Result for Entropy, ARI and F-Measure for MAR and MDA on soybean (small) dataset

Number of cluster	Entropy		Adjusted Rand Index		F-Measure	
	MDA	MAR	MDA	MAR	MDA	MAR
2	1.2090	1.2090	0.2965	0.2965	0.7084	0.7084
3	0.5463	0.5463	0.6537	0.6537	0.8694	0.8684
4	0.5957	0.5002	0.6922	0.7477	0.8784	0.9045

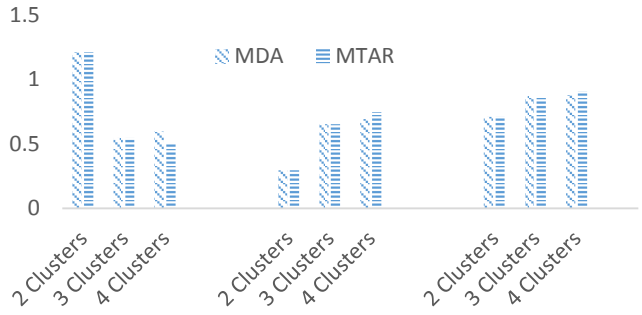

Figure 2: Cluster Measure of MDA and MAR on soybean (small) dataset

For this dataset, as shown on Figure 2, all the validation methods Entropy, ARI and F-Measure shows a similar trend. The validity level seems weaker during the beginning of the clusters formation i.e. when the number of clusters is small. But, when the number of clusters increased, the validity level increasingly improved. For example, in the case of 2 clusters formation, the entropy value for both technique are very high which exceeding 1.0 and it is supported by the value of ARI which is just a bit higher than 0.2. This situation implies that both technique failed to resolve the data uncertainties at this stage due to the limited number of data as well as the small number of clusters. However, F-Measure value for 2 clusters that is higher than 0.6 does not show the same situation, instead, it shows the better validity value for both technique. Next, it can be seen, during the formation of 3 clusters, the validity value is improved as shows by decrement in Entropy value as well as the increment in ARI value, it implies that both techniques MAR and MDA starting to reduce the uncertainties in data as a result of 2 cluster formation. Meanwhile, at this stage, not much differences is observed from the F-Measure results. Finally, the real differences between MDA and MAR is obtains by the result of the final cluster formations i.e. 4 clusters formation. Obviously, from all three methods, MAR technique shows an improvement in validity as compare to MDA. In fact, MDA shows a deterioration which implies that when the number of cluster is greater and finding the differences between the similarity and dissimilarity between objects in cluster are very difficult. But, indirectly it shows that MAR can handle the issue better. The facts is supported by the result of second experiment.

Second experiment used a Zoo dataset which is comprised of one-hundred and one (101) objects of animals. Each animal is then described by the terms of eighteen (18) categorical-valued attributes. The dataset is in the completed state without any missing values. Each animals is classified into one (1) of the seven (7) animal class ranges from one (1) to seven (7). The summary of Entropy, ARI and F-Measure results using Zoo for both techniques is given in Table 4, while the graph that depicting the validity of MAR as compare to MDA is shown in Figure 3.

Table 4
Result for Entropy, ARI and F-Measure for MAR and MDA on zoo dataset

Number of Cluster	Entropy		Adjusted Rand Index		F-Measure	
	MDA	MTAR	MDA	MTAR	MDA	MTAR
2	1.6726	1.6726	0.2510	0.2510	0.7371	0.7371
3	1.2654	1.2654	0.3981	0.3981	0.7905	0.7905
4	1.0963	0.7222	0.4621	0.6811	0.8034	0.8034
5	0.8063	0.1196	0.6595	0.8340	0.8700	0.8824
6	0.6180	0.08972	0.7044	0.8641	0.8727	0.9140
7	0.4626	0.0664	0.8068	0.8520	0.8810	0.8829

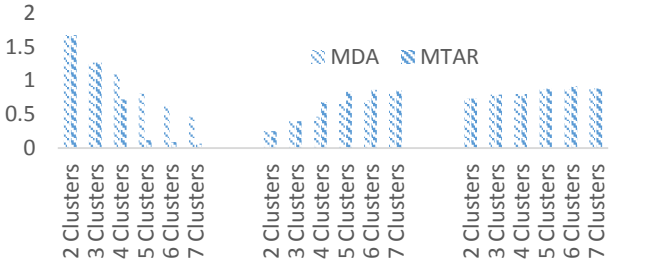

Figure 3: Cluster Measure of MDA and MAR on zoo dataset

It can be seen in Figure 3, all three evaluation methods showed MAR is better as compared to MDA. But, since the differences in value of F-Measure is not very significant, it will not be discussed here. In a case of entropy, a significant differences is shown whereas MAR successfully overcome the MDA almost up to 3 times higher especially when it comes to the formation of 5, 6 and 7 clusters. The same situation also shows by ARI, although the differences is not very significant but it show that MDA still have an issue with uncertainties when involving large dataset and large number of clusters.

Next, to show the performance of the technique in term of cluster validity, a large dataset from Malaysian Ministry of Education is used. The dataset is comprised of 39449 instances of *Ujian Penilaian Sekolah Rendah* (UPSR) results which contains 13 attributes: BMK *(Bahasa Melayu Kefahaman)*, BMP *(Bahasa Malaysia Penulisan)*, BMKC *(Bahasa Melayu Kefahaman untuk aliran Cina)*, BMPC *(Bahasa Melayu Penulisan untuk aliran Cina)*, BMKT *(Bahasa Melayu Kefahaman untuk aliran Tamil)*, BMPT*(Bahasa Melayu Penulisan untuk Aliran Tamil)*, English, Mathematics, Science, BCTK*(Bahasa Cina/Tamil Kefahaman)*, BCTP *(Bahasa Cina/Tamil Penulisan)*, PEKA*(PentksiranKemahiranAmali)* and KAFA*(Kemahiran Agama dan Fardhu Ain)*. If the subjects is not applicable to student, zero (0) is inserted. Eleven (11) attributes has eight (8) categorical variables including zero (0) while PEKA and KAFA respectively have five (5) and four (4) categorical variables. MAR technique is applied to determine the attribute that can be used to cluster the students by using only the information in the dataset. In addition, the UPSR knowledge domain is used to assists the clustering process.

On the first run, MTAR technique have chooses BMK attribute and categorical variable '0' as the partition attribute. By using this result, the dataset is partitioned into two. Further analysis on the results shows that the first partition is a collection of result for SK while second partition is a result

for SJK as shown in Figure 4. It is clear, MAR had determine the correct attribute for partitioning, which is equivalent to current practices. But, based on the knowledge domain, both clusters required a further clustering. Next, the MAR technique is applied on the SJK cluster, where BMKT attributes with categorical value '0' has been chosen as the partition attributes. As shown in Figure 4, BMKT dividing SJK into two clusters: SKJC and SKJT. Obviously, up to this stage, the choice made by MAR is to segregate the students accordingly by their school type. In other word, MAR has just determined that the school type is a one of the discriminants factors.

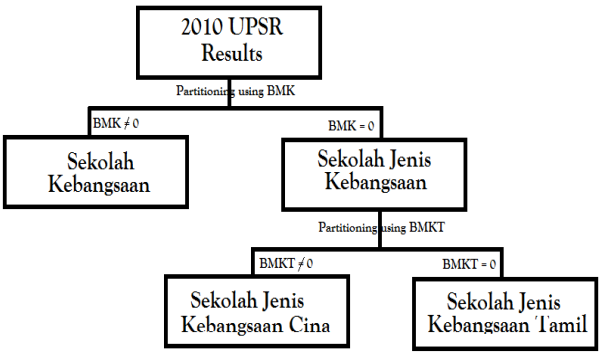

Figure 4: Clustering Result on UPSR dataset

Next, the experiment is continue on the *Sekolah Kebangsaan* partition, which shows that the data was divided into 12 clusters as in Figure 5. It can be seen that, out of twelve clusters, ten clusters are above 70% of in-clusters similarity which indirectly shows that MAR has successfully cluster the big data into their corresponding significant cluster. A further analysis should be carry-out to understand the results.

In summary, it is clear that MAR technique can be used as an alternative tool for categorical data clustering with a better cluster validity. In addition, with the experiment on the real dataset that shows an equivalent results to the current practice, thereby strengthens the capability of MAR technique itself.

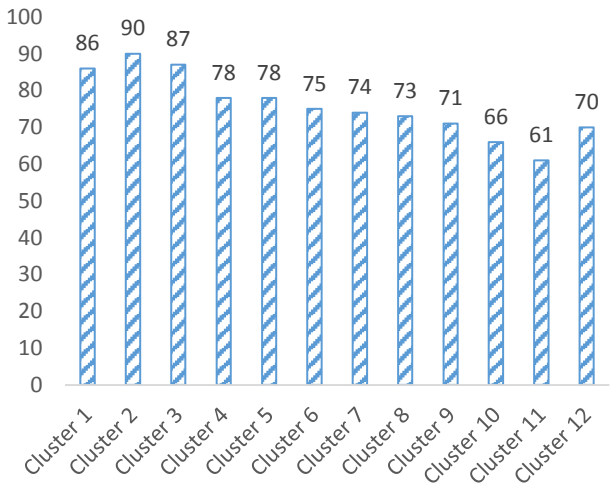

Figure 5: In clusters similarity (%)

VI. Conclusion

Higher cluster validity is a main concern in every clustering technique. Providing higher cluster validity usually involves the understanding on the uncertainty in data. One of the method to organize the uncertainty in data is by using a mathematical tools. In this paper, the validity of the cluster produced by a technique called Maximum Attribute Relative has been presented. A technique which used a new mathematical tool called soft-set theory was evaluated using three methods of external cluster validation measurements: Entropy, Adjusted Rank Index and F-Measure. All three results from the experiments show that MAR technique is produced a cluster with better validity as compared to MDA technique that is based on the rough set theory. Validation using real dataset from Malaysia Ministry of Education also shown some equivalent result as compare to the current practice on the same dataset.

References

[1] D. Molodtsov, "Soft set theory-first results," in *Computer and Mathematics with Applications*, vol. 37, no. 4/5, pp. 19-31, 1999.

[2] R. Mamat, M. M. Deris and T. Herawan, "MAR – Maximum attribute relative of soft-set for partition attribute selection," *Knowledge Based System*, vol. 52, pp. 11-20, 2013.

[3] Z. Pawlak, "Hard and soft sets," in *RSKD '93 Proceedings of the International Workshop on Rough Sets and Knowledge Discovery: Rough Sets, Fuzzy Sets and Knowledge Discovery*, 1993, pp. 130-135.

[4] D. Pei, and D. Miao, "From soft sets to information systems," in *2005 IEEE International Conference on Granular Computing*, vol. 2, 2005, pp. 617-621.

[5] Z. Pawlak, "Rough sets," *International Journal of Computer and Information Sciences*, vol. 11, no. 5, pp. 342-356, 1982.

[6] Z. Pawlak, *Rough Sets: Theoretical Aspect of Reasoning about Data*. Netherlands: Springer, 1991.

[7] Y. Y. Yao, and N. Zhong, "Granular computing using information tables," in *Data Mining, Rough Sets and Granular Computing*, T. Y. Lin, Y. Y. Yao, and L. A. Zadeh, Eds. Heidelberg: Physica, 2002, pp. 102-124.

[8] Z. Pawlak, "Rough set approach to knowledge-based decision support," *European Journal of Operational Research*, vol. 99, no. 1, pp. 48-57. 1997.

[9] Z. Pawlak, and A. Skowron, "Rudiments of rough set," *Information Sciences*, vol. 177, no. 1, pp. 3-27, 2002.

[10] T. Herawan, and M. M. Deris, "On multi-soft sets construction in information systems," in *Emerging Intelligent Computing Technology and Applications with Aspects of Artificial Intelligence*, D.-S. Huang, K.-H. Jo, H.-H. Lee, H.-J. Kang, and V. Bevilacqua, Eds. Berlin, Heidelberg: Springer, 2009, pp. 101-110.

[11] S. C. Sripada, and S. M. Rao, "Comparison of purity and entropy of k-means clustering and c-means clustering," *Indian Journal of Computer Science and Engineering*, vol. 2, no. 3, pp. 343-346. 2011.

[12] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-659, 1948.

[13] H. Xiong, M. Steinbach, P. N. Tan, and V. Kumar, "HICAP: Hierachical clustering with pattern preservation," in *Proceeding of the 4th SIAM International Conference on Data Mining*, 2004, pp. 279-290.

[14] Y. Zhao, G. Karyis, and U. Fayyad, "Hierachical clustering algorithms for document datasets," *Data Mining and Knowledge Discovery*, vol. 10, no. 2, pp. 141-168, 2005.

[15] H. Kim, and H. Park, "Sparse non-negative matrix factorization via alternating non-negative constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495-1502, 2007.

[16] L. Hubert, and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193-218, 1985.

[17] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of American Statistical Association*, vol. 66, no. 36, pp. 846-850, 1971.

[18] K. Y. Yeung, and W. L. Ruzzo, "Principal components analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763-774, 2001.

[19] M. J. Santos, and M. Embrechts, "On the use of the adjusted rand index as a metric for evaluating supervised classification," *Artificial Neural Networks – ICANN 2009*, C. Alippi, M. Polycarpou, C. Panayiotou, and

G. Ellinas, Eds. Berlin, Heidelberg: Springer, vol. 5769, 2009, pp.175-184.

[20] C. J. V. Rijssbergen, "Foundation of evaluation," *Journal of Documentation*, vol. 30, no. 4, pp. 365-373, 1974.

[21] R. Marxer, P. Holonowicz, P. Purwins, and A. H. Hazan, "An f-measure for evaluating of unsupervised clustering with non-determined number of clusters," *Technical Report, UniversitatPempeuFabra*, 2008.

[22] T. Herawan, M. M. Deris, and J. H. Abawajy, "A rough set approach for selecting clustering attribute," *Knowledge Based System*, vol. 23, no. 3, pp. 220-231, 2010.