Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). 7/2 (2014), 61-66
DOI: http://dx.doi.org/10.21609/jiki.v7i2.258

DIVERSITY-BASED ATTRIBUTE WEIGHTING FOR K-MODES CLUSTERING

M Misbachul huda, Dian Rahma Latifa Hayun, and Annisaa Sri I.

Informatics Engineering, information Technology Department
Institut Teknologi Sepuluh Nopember Surabaya, East Java, Indonesia

E-mail: annisaaindrawanti@gmail.com, misbachul@mhs.if.its.ac.id

Abstract

Categorical data is a kind of data that is used for computational in computer science. To obtain the information from categorical data input, it needs a clustering algorithm. There are so many clustering algorithms that are given by the researchers. One of the clustering algorithms for categorical data is k-modes. K-modes uses a simple matching approach. This simple matching approach uses similarity values. In K-modes, the two similar objects have similarity value 1, and 0 if it is otherwise. Actually, in each attribute, there are some kinds of different attribute value and each kind of attribute value has different number. The similarity value 0 and 1 is not enough to represent the real semantic distance between a data object and a cluster. Thus in this paper, we generalize a k-modes algorithm for categorical data by adding the weight and diversity value of each attribute value to optimize categorical data clustering.

Keywords: *categorical data, diversity, K-modes, attribute weighting.*

Abstrak

Data Kategorial merupakan suatu jenis data perhitungan di ilmu komputer .Untuk mendapatkan informasi dari input data kategorial diperlukan algoritma klastering. Ada berbagai jenis algoritma klastering yang dikembangkan peneliti terdahulu. Salah satunya adalah K-modes. K-modes menggunakan pendekatan simple matching. Pendekatan simple matching ini menggunakan nilai similarity. Pada K-modes, jika dua objek data mirip, maka akan diberi nilai. Jika dua objek data tidak mirip, maka diberi nilai 0. Pada kenyataannya, tiap atribut data terdiri dari beberapa jenis nilai atribut dan tiap jenis nilai atribut terdiri dari jumlah yang berbeda. Nilai similarity 0 dan 1 kurang merepresentasi jarak antara sebuah objek data dan klaster secara nyata. Oleh karena itu, pada paper ini, kami mengembangkan algoritma K-modes untuk data kategorial dengan penambahan bobot dan nilai *diversity* pada setiap atribut untuk mengoptimalkan klastering data kategorial.

Kata Kunci: *data kategorial, diversity, K-modes, pembobotan atribut.*

1. Introduction

Computer science is always related to the data. All computational processes need data not only small data but also big data. There are two data types, they are numeric data and categorical data. Arithmetic process can be given to numeric data but it cannot be given to categorical data. Categorical data is used in some systems or applications. For example, categorical data in intrusion detection systems, population data and customer information in online shopping. Example of categorical data in intrusion detection system is IP address. IP address in each data must be different and it cannot be arithmetically compared. It will be in population data and customer information, too. Population data has such categorical data like gender, blood type and home address. Costumer information, in online shopping, has such categorical data. For example: the phone number and the used bank. The categorical data clustering considers the similarity or dissimilarity between data. Similarity or dissimilarity can be considered by distance between the two data object. The shorter the distance between objects, the more similar the objects. Simple matching approaches can be used to calculate the distance. Example of simple matching approach for categorical data is k-modes [7]. In intrusion detection system, there is such new intrusion that has not been known earlier. It needs a clustering algorithm to cluster and detect the new intrusion. Clustering is used to cluster population data and customer data to obtain the information needed, too.

The study [7], k-means and k-modes are joined together to cluster numeric and categorical data. K-means is a clustering algorithm for numeric data. K-means clusters the data type that can

be arithmetically compared. Categorical data is a kind of data that cannot be arithmetically compared each other. K-means cannot be used to cluster categorical data. To overcome the categorical data, it uses k-modes algorithm. K-modes algorithm uses a simple matching approach to the process of matching dissimilarity clustering of data, replacing the means to modes. In k-means, to update a cluster centroid, it is used means formulae. On the other hand the k-modes, modes updating on the clustering process is determined by the frequency of occurrence of data so as to minimize the cost function.

In [3], K-modes algorithm is supposed to have some deficiencies. In K-modes, the two similar objects have similarity value 1, and 0 if it is otherwise. According to [3], a simple matching approach 0 and 1 are not good enough to represent the real semantic distance between a data object and a cluster. Thus, in [3] the authors proposed a range between 0 and 1 which represents the weight of similarity. In [8], focused on optimizing the k-modes algorithm for clustering categorical data with the new dissimilarity approach using the rough set membership. According to the research, the shortcoming of the simple matching approach is having a weak intra-similarity. To solve the issue, the dissimilarity approach (frequency-based) between the two objects in [8] takes into account to the distribution of universal attribute values. The study [1] explains that sometimes the occurrence of low frequency data and high frequency data have the same overlap similarity value. With these problems, the study [1] approach takes into account to the similarity of the frequency distribution of the different attributes value to define the similarity between two categorical attribute values.

In addition, the dissimilarity approach taking into account the frequency distribution of different attribute values can also be applied to the optimization of categorical data dissimilarity [1]. For example, there are two data, AAABB and ABCDD. If using the k-modes algorithm [8], both dissimilarity values are equal to 1. As we can see, the diversity level between the first data and the second data is different because the quantity of each letter is different. Thus, in this paper, the research focuses on the k-modes clustering with diversity-based attribute weighting and the research contribution is distance values diversity to optimize the categorical data dissimilarity.

2. Methods

Categorical Data

Categorical variables represent types of data which may be divided into groups. Analysis of categorical data generally involves the use of data tables. A two-way table presents categorical data by counting the number of observations that fall into each group for two variables, one divided into rows and the other divided into columns.

There is no intrinsic ordering to the categories in categorical data. For example, gender is a categorical variable having two categories (male and female) and there is no intrinsic ordering to the categories. Hair color is also a categorical variable having a number of categories (blonde, brown, brunette, red, etc.) and again, there is no agreed way to order these from highest to lowest.

So that, computing the similarity and the dissimilarity between categorical data instances is not straightforward owing to the fact that there is no explicit notion of ordering between categorical values. To tackle this, there are some data-driven similarity measures that have been proposed for categorical data. In this section, we will describe the categorical data characteristics.

But first, we will define the notation used in later explanation. Let's consider for a categorical data set *D* containing N objects, defined over a set of l attribute where A_i denotes the i^{th} *attribute.* And let the n_i be the number of values in each attributeA_i. The next notation used are following:
$f_i(x) =$ the frequency of value x lies on attribute A_i in data set D.
$p_i(x) =$ the probability of attribute A_i to take the value x in the data set D, defined as equation(1).

$$p_i(x) = \frac{f_i(x)}{N} \quad (1)$$

$p_i^2(x) =$ another probability measure of attribute A_i to take the value x in the data set D, defined as equation(2).

$$p_i^2(x) = \frac{f_i(x)(f_i(x)-1)}{N(N-1)} \quad (2)$$

Size of the data, N. Most measure are typically invariant to the size of the data. *Number of attributes, l.* In [1] experiment showed that the number of attributes does affect the performance of the outlier detection algorithm.

The frequency of values taken by each attribute, n_i. A data set may contain attributes that take some values and attributes that only take a few values. A similarity measure may ignore another attribute while finding the more importance attribute.

Distribution of $f_i(x)$. It refers to the distribution of frequency of values taken by attribute in the given data set. It is possible for a similarity measure to give more priority to frequently occurring attribute values or otherwise.

Similarity and Dissimilarity Measure for Categorical Data in K-Modes

The classical theory of clustering in k-Modes, either an element belongs to a cluster or it does not. The corresponding membership function is the characteristic function of the cluster that takes values 1 and 0. Suppose that there are $x,y \in D$, then the simple matching dissimilarity measure in k-Modes is defined following equation(3).

$$Dis(x,y) = \begin{cases} 1, if\ x \neq y \\ 0, otherwise \end{cases} \quad (3)$$

With this concept, the distance between two objects computed often results in cluster with weak intra-similarity and disregards the similarity embedded in the categorical values [1].

Then, a new dissimilarity measure between the mode of a cluster and an object is introduced for the k-Modes Algorithm. To obtain such a cluster having the strong intra-similarity, the rough membership is used. It takes values between 0 and 1.

Taking account of the frequency of mode components in the current cluster, Ng et al. [2] introduced a valuable dissimilarity measure into the *k*-Modes clustering algorithm. Let $P \subseteq A$ and $a \in P$, then the Ng dissimilarity measure $Dis_P(z_l, x_i)$ between a categorical object x_i and the mode of a cluster z_l with respect to P is defined as the following equation(4).

$$Dis_P(z_l, x_i) = \sum_{a \in P} Dis_a(z_l, x_i),$$

where,

$$Dis_a(z_l, x_i) = \begin{cases} 1, if\ f(z_l, a) \neq f(x_i, a), \\ 1 - m_a, otherwise \end{cases}$$

where,

$$m_a = \frac{|\{x_i | f(x_i, a) = f(z_l, a), x_i \in c_l\}|}{|c_l|} \quad (4)$$

And $|c_l|$ is the number of objects in the *lth* cluster. For the k-Modes algorithm with Ng's dissimilarity measure [2], the simple matching dissimilarity measure is still used in the first iteration. So, Cao et al. in [3] proposed a new dissimilarity measure by using $Sim_a(x,y)$ defined in equation (5).

$$Sim_a(x,y) = \frac{f(x,a) \equiv f(y,a)}{\sum_{z \in D} f(x,a) \equiv f(z,a)} \quad (5)$$

From the equation(5) formula [3] introduced that the new dissimilarity measure is defined as the following equation(6).

$$NDis_P(z_l, x_i) = \sum_{a \in P} NDis_a(z_l, x_i) \quad (6)$$

where,

$$NDis_a(z_l, x_i) = 1 - Sim_a(z_l, x_i) \times m_a$$

As opposed to Ng's dissimilarity measure, the similarity $Sim_a(z_l, x_i)$ between object $x_i$and cluster z_l is included in the proposed measure $NDis_a(z_l, x_i)$. In [3] introduced a weighted k-modes clustering algorithm, by considering that the similarity value between two objects is not always 1, but it can be a value between 0 and 1. A pair of object x_i, x_j is considered more similar than the second pair(x_s, x_t), if and only if $x_i$and x_j exhibit a less common attribute value match in the population. In other words, similarity among objects could be decided by the un-commonality of their attribute value matches. Based on that, in [3] defined the "More Similar Attribute Set" of an attribute value $a_j^{(r)}$ as the following equation(7).

$$M(a_j^{(r)}) = \{a_j^{(t)} | f(a_j^{(t)}|D) \leq f(a_j^{(r)}|D)\} \quad (7)$$

where $f(a_j^{(t)}|D)$ is the frequency count of attribute $a_j^{(t)}$ in the data set D. This is the set of attribute values with lower or equal frequencies of occurrence than that of $a_j^{(r)}$. Note that a value pair is more similar if it has lower frequency of occurrence. The weighting function in [3] is defined as equation(8).

$$\omega(a_j^{(r)}) = 1 - \sum_{a_j^{(t)} \in M(a_j^{(r)})} \frac{f(a_j^{(t)}|D)(f(a_j^{(t)}|D) - 1)}{n(n-1)} \quad (8)$$

where n is the frequency of objects in the data set D. If above function is used, less frequent values will make more contributions to the similarity value.

Diversity Index and Variable Uniqueness

Although the above algorithms can effectively improve the accuracy of the clustering result of the k-modes algorithm, it is noted that the k-modes algorithm and its modified version cannot detect the diversity of the data in a certain attribute in data set. It is easily prove in simple matching similarity measure. Because in the value will be 1 if it is identical, and 0 if otherwise. In wk-modes, the basic

TABLE 1
DATA SAMPLE

Dataset/ object	x_1	x_2	x_3	x_4	x_5	Div
D_1	A	A	A	A	A	1
D_2	A	A	A	A	B	2
D_3	A	A	A	B	B	2
D_4	A	A	B	B	C	3
D_5	A	A	B	C	D	4
D_6	A	B	C	C	C	3
D_7	A	B	C	D	E	5

concept used is using individual Simpson diversity index (will be explained in the next part). By using this information, it is noted that this algorithm cannot detect the diversity of data in data set. Let us see Table 1 as the data sample of counting the dissimilarity between two objects. Let the number of attribute in the given data set is only 1, and the number of data set is 7. The number of object in a dataset is 5. From the table we have the diversity value (Div) as variable uniqueness shown in the table.

The value of Div show us the number of distinct value in a dataset. We assume that by using this Div value it will increase the dissimilarity between clusters based on the diversity of attribute value.

From the Table 1, we will get 1 as the value of dissimilarity because each data is different. But, as we can see in the real in each data, there are some similar data in it. But, they have different diversity. For AAAAA, the diversity value is 1 because there's just a single object data. But for ABCDE, the diversity value is 5 because actually there are 5 different data. Based on this diversity, we propose a new method to count the dissimilarity measure to count the distance between two distinct objects. We use this diversity value to decrease the inter-similar cluster value.

Diversity index is formerly a concept in biological area. It is a mathematical measure of species diversity in a community. Diversity indices provide more information about community composition than simply species richness (i.e., the number of species present); they also take the relative abundances of different species into account [4]. Diversity indices provide important information about rarity and commonness of species in a community. The ability to quantify diversity in this way is an important tool for biologists trying to understand community structure.

One of the commonly used diversity index measures is Simpson Diversity Index (SDI). The basic concept of SDI represents the measurement of dissimilarity by frequency based. Simpson's Diversity Index is a measure of diversity. In ecology, it is often used to quantify the biodiversity of a habitat. It takes into account the number of species present, as well as the abundance of each species. Simpson's Index **(**D**)** measures the probability that two individuals randomly selected from a sample will belong to the same species (or some category other than species). The formula is defined as below.

$$D = \frac{\sum n(n-1)}{N(N-1)} \tag{9}$$

where n is the total number of organisms of a particular species, and N is the total number of organisms of all species. Simpson's Index gives more weight to the more abundant species in a sample. The addition of rare species to a sample causes only small changes in the value of D [5]. In clustering theory, the property n can be defined as the number of object x_i in a dataset D, and the property N can be defined as the number of all objects in data set D.

Diversity-based Attribute Weighting for K-Modes Clustering

Taking into account of the intra and inter cluster information, we introduce the new dissimilarity measure in the equation(10).

$$Dis_P(x_{i,}y_j) = \frac{SDI_i \times Div \times n_i}{N} \times \frac{SDI_j \times Div \times n_j}{N} \tag{10}$$

where, SDI_i is simpson diversity index in data set D for object $x_{i,}$ SDI_j is simpson diversity index in data set D for object y_j, Div is the number of category in data set D, n_i is the frequency of x_i in a cluster, n_j the frequency of y_j in a cluster, N is the number of data in data set D.

This formula is derived from the basic concept of probability count of object x_i as defined below.

$$p(x_i) = \frac{N_{x_i}}{N} \tag{11}$$

where, $p(x_{i,})$ is the probability of x_i, Nxi is the frequency of x_i, and N is the number of data in data set D.

In order to count the dissimilarity between two objects $x_{i,}y_j$, we modify the above formula as defined in equation(12).

$$Dis_P(x_i, y_j) = \frac{N_{x_i}}{N} \times \frac{N_{y_j}}{N} \tag{12}$$

TABLE 2
SCALABILITY EXPERIMENTAL RESULT

Pengujian Scalability		
Number of objects	Diversity based Time(ms)	Original Time (ms)
10	4	3
50	10	10
100	18	18
500	135	124
1000	737	585
5000	38010	18984

TABLE 3.
DATA SET CHARACTERISTIC

Dataset	Class number	Number of data	Number of attributes
Voting	2	435	16
Mushroom	2	8124	23
Soybean	4	47	36

where, $Dis_P(x_i, y_j)$is the dissimilarity of x_i, y_jin attribute P, $N_{x_i}$is the number of object $x_i$in data set D, $N_{y_j}$is the number of object $y_j$in data set D and $N$is the number of data in data set D.

To increase the accuracy of clustering process, we add the Simpson diversity index that has defined in the previous to the formula above. Our new proposing dissimilarity measure is defined as the following equation(13).

$$Dis_P(x_{i,}y_j) = \frac{SDI_i \times Div \times n_i}{N} \times \frac{SDI_j \times Div \times n_j}{N} \quad (13)$$

where, $SDI_i$is simpson diversity index in data set D for object $x_{i,}$ $SDI_j$is simpson diversity index in data set D for object y_j, $Div$is the number of category in data set D, $N$is the number of data in data set D.

3. Results and Analysis

In this part, we will discuss about the experimental result of scalability and efficiency of clustering algorithms. It employs three different algorithms. In first part of this section, we will explain the experiment environment and evaluation index. In second part, we will explain the experimental result of scalability of original k-modes and our proposed method. And in the last part, we will show the experimental result of clustering efficiency of four different modified k-modes algorithm include our proposed method.

TABLE 4
CLUSTERING EFFICIENCY RESULT

	Voting	Mushroom	Soybean	Average
Original K-Modes	0.8592	0.7381	0.8177	0.805
wk-Modes	0.8651	0.7905	0.897	0.850867
Zengyou k-modes	0.8734	0.7644	0.86	0.8326
Diversity k-modes	0.8861	0.7849	0.887	0.852667

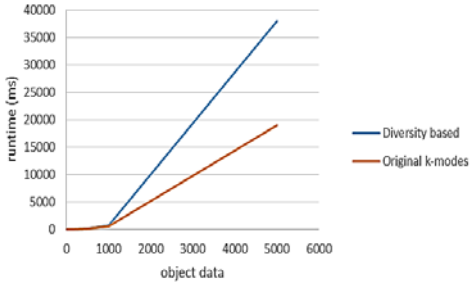

Figure 1. Scalability Graphic between original k-modes and diversity-based k-modes

Experiment Environment and Evaluation Index

The experiment was done in core i3 (1.4 GHz), 4GB RAM, and windows 8.1x64 based computer. The implementation of the algorithm uses Java as the programming language. To evaluate the accuracy of the algorithm, we use the following equation(14).

$$acc = \frac{\sum_{i=1}^{k} a_i}{n} \quad (14)$$

where k is the number of known categories, a_i is the number of object that lies in the right cluster in $C_i (1 \le i \le k)$.

Scalability Evaluation

To compare the scalability of original k-modes and our proposed method, we use synthetic data with the variant number of objects between 10 and 5000. This synthetic data have 23 different attributes. To decrease the random effect of k-modes algorithm, we did 100 times experiment for every scenario.

For each experimental result shown in Table 2, is the average value of the 100 times experimental result. From these experiments, we get a scalability graphic shown in Figure 1.

The experimental result of runtime from diversity-based k-modes show that it needs longer time than the original k-modes to evaluate the gi-

ven data set. It is caused by added step to compute the diversity of the data set.

Clustering Efficiency Evaluation

In this section, to compare the efficiency clustering algorithms, we use four different modified k-modes algorithm include our proposed method. They are Simple-Matching k-modes, Zengyou's k-modes, wk-Modes and our new proposed method diversity-based k-modes. We use tree different data set from UCI Machine Learning [6]. Table 3 shows the characteristic of the data set.

In experiments, the missing value is ignored. To decrease the random effect in k-modes, every experiment was conducted 100 times. Table 4 shows the result of clustering efficiency evaluation between the four algorithms. Every value is the average of 100 times experiments result.

4. Conclusion

The k-modes algorithm is widely used for clustering categorical data. Dissimilarity and similarity measure play crucial rules in this area. In this paper the limitation of the former algorithm on the usage of between cluster information is solved by using simpson diversity index to extend the value of the intra similarity index. The experimental result shows that our proposed algorithm give better result.

References

[1] Shyam Boriah, Varun Chandola, Vipin Kumar. *Similarity Measures for Categorical Data: A Comparative Evaluation*. Department of Computer Science and Engineering University of Minnesota. 2008s

[2] M.K. Ng, M.J. Li, Z.X. Huang, Z.Y. He, *on the impact of dissimilarity measure in k-Modes clustering algorithm,* IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (3) (2007) 503–507.

[3] Zengyou He, Xiaofei Xu, Shengchun Deng. *Attribute value weighting in k-modes clustering*. Elsevier. 2011

[4] M. Beals, L. Gross, and S. Harrell, http://www.tiem.utk.edu/~gross/bioed/bealsmodules/shannonDI.html, 2000, retrieved June 1, 2014

[5] Offwell Woodland & Wildlife Trust http://www.countrysideinfo.co.uk/simpsons.htm., 2000, retrieved June 1, 2014

[6] UCI Machine Learning Repository http://www.ics.uci.edu/mlearn/MLRepository.html, 2009, retrieved June 1, 2014.

[7] Huang, Z.J. Extension to the k-means algorithm for clustering large data sets with categorical values. Data mining Knowledge Discovery, Vol. 2, No. 3, pp.283-304. 1998.

[8] Fuyuan Cao, Jiye Liang, Deyu Li, Liang Bai, Chuangyin Dang. A dissimilarity measure for the k-Modes clustering algorithm. Elsevier. 2012.