International Journal of Database Theory and Application
Vol.9, No.3 (2016), pp. 191-202
http://dx.doi.org/10.14257/ijdta.2016.9.3.19

The Similarity for Nominal Variables Based on F-Divergence

Zhao Liang[*,1] and Liu Jianhui[2]

[1]*Institute of Graduate, Liaoning Technical University, Fuxin, Liaoning, 123000, P.R. China*
[2]*School of Electronic and Information Engineering, Liaoning Technical University, Huludao, Liaoning, 125000, P.R. China*
[1]*ttimefighter@163.comE-mail* [2]*liu_jianhui@outlook.com*

Abstract

Measuring the similarity between nominal variables is an important problem in data mining. It's the base to measure the similarity of data objects which contain nominal variables. There are two kinds of traditional methods for this task, the first one simply distinguish variables by same or not same while the second one measures the similarity based on co-occurrence with variables of other attributes. Though they perform well in some conditions, but are still not enough in accuracy. This paper proposes an algorithm to measure the similarity between nominal variables of the same attribute based on the fact that the similarity between nominal variables depends on the relationship between subsets which hold them in the same dataset. This algorithm use the difference of the distribution which is quantified by f-divergence to form feature vector of nominal variables. The theoretical analysis helps to choose the best metric from four most common used forms of f-divergence. Time complexity of the method is linear with the size of dataset and it makes this method suitable for processing the large-scale data. The experiments which use the derived similarity metrics with K-modes on extensive UCI datasets demonstrate the effectiveness of our proposed method.

Keywords*: Similarity; Nominal variables; f-divergence; K-modes*

1. Introduction

Measuring the similarity between data objects is one of the most important problem in the data mining tasks which involve similarity or distance computation such as clustering, outlier analysis, and nearest-neighbor classification, we need ways to assess how alike or unalike objects are in comparison to one another[1].

If the data objects are defined by the vectors formed with continued variables, the similarity can be computed using Minkowski Distance and the most common ones of them are Manhattan Distance and Euclidean Distance. When the data objects are represented by nominal variables, the similarity cannot be measured straightforwardly because the comparison of categorical variables has only two states of same and not same. Traditional similarity measuring algorithm for categorical variables can be mainly divided into two categories. The distinction between these two categories is whether consider the difference among variables or not.

The representative of the first kind of algorithms is Simple Matching Distance (SMD) [4-5], which is the most common algorithm to measure nominal variables for unsupervised learning. Let X and Y be two data objects described by categorical variables, then the similarity measure between X and Y can be defined by the total mismatches of the corresponding attribute of the two objects. The smaller the number of mismatches is, the more similar the two objects.

However, SMD is too simple to keep much information in datasets, which made researchers to find an appropriate way to measure categorical attributes with data-driven

ISSN: 2005-4270 IJDTA

method. This kind of measures account the frequency distribution of different variables in the same attribute as the key characteristic [4]. Some other algorithms further computing similarity for categorical variables in unsupervised learning are based on frequently co-occurring items [4,7].When under the condition of supervised learning, the relationship between attributes and class label are involved to improve the accuracy.

Stanfill and Waltz proposed VDM(Value Distance Matrix) [2] and then Cost and Salzberg [3] proposed MVDM(Modified Value Distance Matrix) based on VDM. MVDM suggested computing the similarity between two categorical variables with respect to class label. Ahmad and Day [6] proposed a rapid algorithm of MVDM and which considers the relationship between variables and all other attributes columns and it can be used in unsupervised learning. Ganti *et al*. [7] described a notion that to attribute pairs in the same attribute column, co-occurrence with other attribute can show their similarity. Many other similarity measuring methods are also based on frequency of co-occurrence [9-11]. Wang presented the CAVS [8] method which uses a similar formula and considers both intra-coupled and inter-coupled relationship between variables. The second kind of methods based on a theory that greater similarity is assigned to the attribute value pair which owns approximately equal frequencies [7] and furthermore the attribute values are similar if they co-occur with the same relative frequency for other attribute columns.

However, the algorithms mentioned above suffer the different problems, which are showed in the following two examples. Table 1 shows a part data of the UCI dataset Balance. Data objects in the dataset are described by 5 attributes: left-weight, left-distance, right-weight, right-distance and balance status. The first four attributes seem like numerical, but they are treated as categorical variables in computing process. Meanwhile the attribute type of the dataset is also labeled as categorical in UCI machine learning repository. When we measure the similarity between data objects with SMD, instances C_1 and C_2 are 0.25, which equals to the similarity between instances C_1 and C_3. Because the reason is that C_1 and C_2 are both in the R set while C_1 and C_3 belong to the different set R and B, and thus the result is not accurate. Another observation of SMD is similarity between instances C_4 and C_5, *i.e*. 0, which means there is no relationship between them. For they both belong to the subset L, the result is obviously incorrect.

Table 1. A Part Data of the Balance Dataset

	left-weight	left-distance	right-weight	right-distance	balance status
C_1	1	2	1	4	R
C_2	2	1	2	4	R
C_3	3	2	2	3	B
C_4	4	3	5	2	L
C_5	5	4	4	1	L
C_6	3	5	3	5	B

This example shows that treating categorical variables with methods like SMD will take adverse effects in data mining, because of nominal variables carrying more complex information than “same and not same”. But algorithms like SMD neglect the difference among different variables. Since the similarity between categorical variables can not be measured directly, it’s necessary to have the help of the hidden information in the data set for the similarity analysis.

The basic idea of MVDM is that the similar attribute variables in the same column must have the approximately frequency of occurrence. However this assumption is

generally not reasonable. This example is extracted from the Adult dataset in UCI machine learning repository which contains 14 attributes and 2 classes. The probability distribution of attribute column work-class is showed in Figure 1.

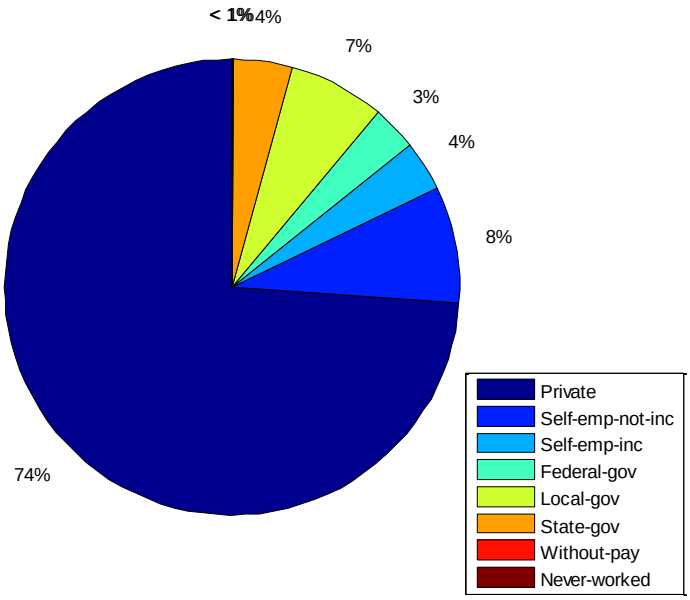

Figure 1. Distribution of Attribute Work-Class

According to the theory that similar variables have approximately frequency, the variable 'Private' will be quite different to all the other variables but as the common sense the work-class of private is not a strange type of work. From the similarity between these values given by MVDM showed in Figure 2, we can see another problem is the almost equal similarities between 'Private' and the other variables which should be different. That means the following step of clustering or classification cannot get useful information. Both examples show that it is very difficult to analyze the similarity for categorical variables.

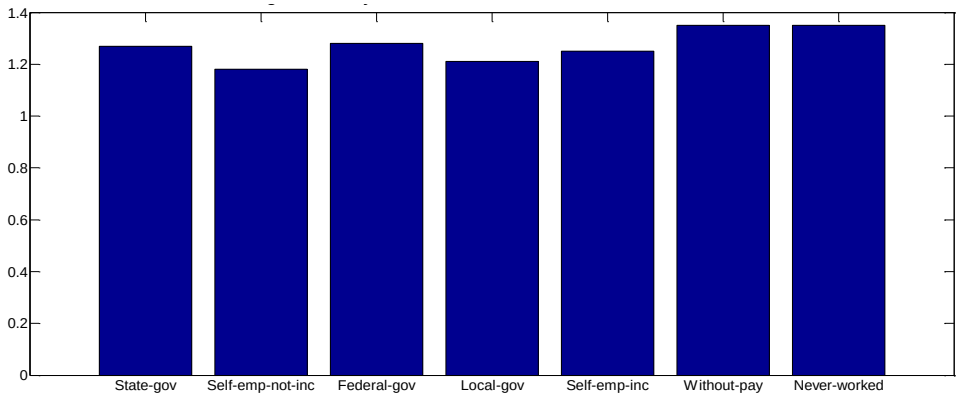

Figure 2. Similarity between 'Private' and other Variables

In this paper we propose a method based on Hellinger Distance to measure the similarity between two categorical variables in unsupervised learning. The algorithm captures the attribute value frequency distribution of different subset separated by categorical variables binding the same attribute and considers the relationship between all attributes in datasets with a high accuracy and relatively low time complexity. We

evaluate theoretical and experimental analysis among the four common forms of f-devergence and choose Hellinger Distance as the optimal one. We compare our proposed measure with two existing metric on extensive UCI categorical data sets in terms of clustering qualities.

Section 2 describes our proposed method and the content in Section 3 is experiments with real world datasets. The result will be compared with other two standard algorithms. Finally we conclude this paper in Section4.

2. Proposed Algorithm

2.1 Definition of Similarity

Similarity can be seen as distance in a measurable space, so if we present the similarity in form of a vector then we can easily compute the distance. Clearly the problem is how to describe the variables with the form of vector. The SMD method focus on measuring the similarity between data objects. The feature vector is formed by only 0 and 1, the similarity is the L_1 norm of the feature vector. The MVDM method computes distance between two categorical variables with respect to class column and uses the difference of the distribution of classes as the characteristic value. The similarity is the L_1 norm of the vector. The unsupervised learning version uses the similarity of variables by all the other attributes as the characteristic values.

Our solution spreads dataset into subsets by categorical variables which we want to measure the similarity between them and every subset presents a variable. In subsets, each dimensionality is presented by a set of distribution. The L_1 norm of the vector which conformed by distances between distributions from different subsets will be the distance between the subsets and this distance can be used as the similarity between two categorical variables by which the two subsets be separated.

2.2 Hellinger Distance

We choose Hellinger distance as the divergence for our method to measure the distance between distributions. In probability and statistics, the Hellinger distance (also called Bhattacharyya distance as this was originally introduced by Anil Kumar Bhattacharya) is used to quantify the similarity between two probability distributions. It is a type of the f-divergence and the f-divergence has many special cases including but not only KL-divergence, χ^2 -divergence and Hellinger distance. The Hellinger distance is defined in terms of the Hellinger integral which is introduced by Ernst Hellinger in 1909 [8-10]. The reason of why we choose the Hellinger distance for our method to measure the distance between distributions will be discussed in Section 3.

Let (Θ, λ) denotes a measurable space with P and Q as two continuous distributions with respect to the parameter λ . The definition of Hellinger distance can be given as

$$D_H(P,Q) = \sqrt{\int_W (\sqrt{P} - \sqrt{Q})^2 d\lambda} \quad (1)$$

It can also be defined for a countable space Φ ,

$$D_H(P,Q) = \sqrt{0.5 * \sum_{\phi \in \Phi} (\sqrt{P(\phi)} - \sqrt{Q(\phi)})^2} \quad (2)$$

2.3 Distance in Unsupervised Learning

We compute distance between two categorical values with respect to every other attribute in dataset as the characteristic value of the distance vector, the L_1 norm will give

the final distance. The proposed algorithm to compute the distance between every pair of categorical variables for all attributes in following manner.

Algorithm HDS (Hellinger Distance Similarity)

Input: Categorical Dataset *D* with *m* attributes and *n* data objects

Output: Distance between all pairs of attribute variables for all attributes

1:For each attribute A_i {

2: Separate dataset into subsets by categorical variables in A_i

3: For each pair of subsets (w_1, w_2) {

4: For each other Attribute A_j ($A_j \neq A_i$){

5: Compute Hellinger distance

6: $D_H(w_1, w_2, A_j) = \sqrt{0.5 * \sum_{\phi \in A_j} (\sqrt{w_1(\phi)} - \sqrt{w_2(\phi)})^2}$

7: }

8: Compute L_1 norm for distance vector $d_H(w_1, w_2) = \sum_{j=1, j \neq i}^{m} d_H(w_1, w_2, A_j)$

9: }

10:}

For example, as the dataset shown in Table 1, the categorical variables R, L, B separate the dataset into 3 subsets. The subset R conformed by data objects C_1 and C_2 which has 4 distributions, the distribution of attribute left-weight is showed in Table 2. The subset L is conformed by data objects C_4 and C_5, which distributions of attribute left-weight is showed in Table3. The distance for left-weight attribute between these two subset computed by 2.1 is

$$D_H(\mathrm{R}, \mathrm{L}, \mathrm{left-weight})^2 = 0.5 * ((\sqrt{0.5} - \sqrt{0})^2 + (\sqrt{0.5} - \sqrt{0})^2 + (\sqrt{0} - \sqrt{0})^2 + (\sqrt{0} - \sqrt{0.5})^2 + (\sqrt{0} - \sqrt{0.5})^2) = 1$$

The distances for 3 other attributes can also be computed by the same formula. For this dataset, they all equal to 1.

$$D_H(R, L) = 1 + 1 + 1 + 1 = 4$$

High value of the distance suggests the high dissimilarity level between the subsets. Obviously, if every pair of distributions is similar, the distance between the subset will be close to 0.

Table 2. Distributions of Attribute Left-Weight in Subset R

X	1	2	3	4	5
P_x	0.5	0.5	0	0	0

Table 3. Distributions of Subset L

X	1	2	3	4	5
P_x	0	0	0	0.5	0.5

Because the categorical variables can not be measured directly like the numerical variables, the information hidden in the dataset can be used for measuring. The MVDM and algorithms similar to it use the co-occurrence as the material of the similarity. However, this method dose not fit all conditions. Thus, we take the distribution as the characteristic and use Hellinger Distance to measure the dissimilarity between the subset instead.

3. Theoretical Analysis

3.1 Why Hellinger Distance

This section compares four common used f-divergence to explain why the Hellinger Distance is finally be used in our algorithm.

In mathematics, a metric or a distance function is required to satisfy the following conditions: non-negativity, symmetry, and triangle inequality. If only satisfy the first two conditions are satisfied, the functions will be called semi-metric. Some functions can only satisfy the condition of non-negativity then they will be called divergence. Machine learning usually use the f-divergence to measure the probability.

Let $f(t)$ be a convex function defined for t>0, with $f(1)=0$. The f-divergence between two probability distributions P and Q is defined by[11]

$$d_{kl}(P \| Q) = \sum_{a} Q(x) f\left(\frac{P(x)}{Q(x)}\right) \tag{3}$$

The f-divergences is introduced and studied independently by Csiszar [12], Morimoto [13] and Ali & Silvey [14] and is sometimes known as Csiszar f-divergence, Csiszar-Morimoto divergence or Ali-Silvey distance. The most common examples of f-divergence are Kullback-Liebler divergence, Jensen-Shannon divergence, Pearson-x^2 divergence and Hellinger Distance.

When $f(t) = t \ln t \Rightarrow d_f(P \| Q) = \sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{q_i}\right)$, the case is Kullback-Liebler divergence which is also called relative entropy, cross entropy or directed divergence. The KL-divergence is non-negative but it does not has the symmetric property and also does not obey the Triangle inequality. In additional, the KL-divergence has another disadvantage: if $q_i=0$ and $p_i/=0$, the divergence is meaningless. Another popular method of measuring the similarity between two probability distributions is Jensen-Shannon divergence. It is based on the KL-divergence.

$$d_{js}(P \| Q) = \frac{1}{2}(d_{kl}(P \| Q) + d_{kl}(Q \| P)) = 0.5 * \sum_{i=1}^{n} (p_i - q_i) \ln\left(\frac{p_i}{q_i}\right) \tag{4}$$

The Jensen-Shannon divergence is symmetric, nonnegative and obeys the Triangle inequality. To avoid the disadvantage like KL-divergence, an improvement of Jensen-Shannon divergence is Jensen Shannon Distance. The Jensen-Shannon Distance is defined

$$D_{js}(P \| Q) = \frac{1}{2}(d_{kl}(P \| M) + d_{kl}(Q \| M)) = 0.5 * \left(\sum_{i=1}^{n} p_i \ln\left(\frac{2p_i}{p_i + q_i}\right) + \sum_{i=1}^{n} q_i \ln\left(\frac{2q_i}{p_i + q_i}\right)\right) \tag{5}$$

When $f(t) = (t-1)^2 \Rightarrow d_f(P \| Q) = \sum_{a} \frac{(P(a) - Q(a))^2}{Q(a)}$, the case is Pearson-$x^2$ divergence. The Pearson-x^2 divergence requests two sets of complete finite discrete probability distributions P and Q which meet the conditions of $p_i \geq 0, \sum_{i=1}^{n} p_i = 1; q_i \geq 0, \sum_{i=1}^{n} q_i = 1$.Similarly, the Person-$x^2$ has the non-negative property but is non-negative and does not obey the Triangle inequality.

When $f(t) = 1 - \sqrt{t} \Rightarrow D_f(P \| Q) = 1 - \sum_a \sqrt{P(a)Q(a)}$, the case is Hellinger Distance. The Hellinger Distance has the symmetric, non-negative properties and obeys the Triangle inequality.

Table 4. Properties of Divergences

divergence	Non-negativity	symmetry	Triangle inequality
KL-divergence	T	F	F
Jensen-Shannon Distance	T	T	T
Pearson-x^2 divergence	T	F	F
Hellinger Distance	T	T	T

Table 4 shows the conditions of 4 commonly used f- divergences introduced above. It's easily to see that Jensen-Shannon Distance and Hellinger Distance satisfied more conditions than the others. That means the Jensen-Shannon Distance and Hellinger Distance are more qualified the algorithms based on distance.

Table 5. Time Consumptions of Divergences

Divergences	KL	JS	Pearson	Hellinger
1st	0.066504	0.116691	0.029401	0.041866
2nd	0.078746	0.106554	0.021279	0.034578
3rd	0.07099	0.126986	0.022231	0.044812
4th	0.063770	0.113847	0.026169	0.037332
mean	0.0700025	0.1160195	0.02477	0.039647

Because all the divergences have the same calculation steps, the difference of computational complexities in applications depend on f(t). Table 7 shows the time consumptions took by 4 divergences for a same random data set which formed by 100000 rows and 2 columns for 4 times. The row of mean show the means of result of all 4 times experiments. Obviously the Pearson-x^2 is the most efficient divergence, then the Hellinger Distance, KL-divergence and Jenson-Shannon Distance. Considering both the properties of metric and complexity comparison, the Hellinger Distance has the best performance. That is why we choose it for our proposed algorithm.

3.2 Complexity of the Algorithm

Each time we compute the Hellinger Distance between two distributions we need to read two columns, one column contains the variables by which we separate the dataset the other belongs to the rest of the attribute. Let the dataset has *m* attribute columns and *n* data objects, the maximum number of attribute values in a single attribute column is *a*. When we computing the Hellinger Distance with two attribute columns, computing the distributions will take m*n steps and computing f(t) for summation will take m*a*(a-1)/2 steps at most. So the upper bound of complexity of the algorithm for the whole dataset will be O ($m^2n+m^2a^2$). This shows that our algorithm is linear with respect to number of data objects in the dataset.

4. Experiments

We perform several experiments on extensive UCI datasets in this section to find out the effectiveness and efficiency of our proposed similarity. The k-means algorithm is the most common partitioning clustering method and only fits the numerical datasets. The k-modes algorithm [1] is the extension of the k-means algorithm which replace the means of clusters with modes and use the simple matching similarity to deal with categorical variables. The k-modes algorithm divides the dataset into k clusters by minimizing the cost function $f_c = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, C_i)$, where C_i is the center of ith cluster. We apply our similarity into k-modes clustering algorithm, analyze quality of clusters and accuracy of clustering.

There are 2 kinds of methods to evaluate the clustering algorithm: the intrinsic and extrinsic methods. The intrinsic methods can be used without the ground truth of the dataset. They evaluate a clustering by examining how well the clusters are separated and how compact the clusters are. But usually this kind of methods need to define a unified dispersion measure and cluster similarity measure, it is not meet the requirement of evaluating different similarities. The extrinsic methods compare the ground truth with a clustering to assess the clustering. We can commonly use the accuracy of a classification dataset to compare different clustering results.

4.1 Intrinsic Method

Some researcher [6] consider that the parameter like DB index can be used to evaluate different similarity in cluster algorithm after the similarities are normalized, but we do not think so. First we introduce a validity index of intrinsic methods. The Davies-Bouldin (DB) index [15] is a validity index introduced by David L. Davies and Donald W. Bouldin for evaluating clustering algorithms. This is an internal evaluation metric of how well the clustering has been done. The DB index is defined as $V_{DB} = \frac{1}{k} \sum_{i=1}^{k} R_i$, where k is the number of clusters and R_i is defined as $R_i = \max_{j \neq i} R_{ij}$. $R_{ij} = \frac{S_i + S_j}{D_{ij}}$, $D_{ij} = d(v_i, v_j)$, where v_i, v_j are the centroids of clusters C_i and C_j. The dispersion measure S of a cluster C is defined as $S_i = \frac{1}{|C_i|} \sum_{x \in C_i} d(x, c_i)$, $|C_i|$ is the number of data points in cluster C_i, c_i is the center or representative data point of cluster C_i and $d(x, c_i)$ is the distance between x and c_i.

It's easily to find out that S presents how compact the clusters are and D presents how well the clusters are separated. The normalized distance between two data objects is defined $D_n = d_o / d_{max}$, where D_n is the normalized distance, d_o is the distance computed by a certain similarity and d_{max} is the maximum distance computed by the certain similarity. Normalization bring the distance computed by different similarity between 0 and 1 and make them can be compared on same scale. The range of Hellinger Distance and simple matching similarity are both [0, 1], so when we compare these two similarity, they can be treat as normalized already.

However, we still cannot use these two parameters in our compare of different similarity. Because the simple matching similarity based on the principle that the distance is 0 with the identical values and is otherwise 1, the other normalized similarity take the distance 0 with identical values but the distance otherwise is maybe in the range of (0,1). So though the ranges of similarities are same, the distances between the same pair of data objects computed by simple matching similarity is not less than the distance computed by other similarities. The experiment result given by [6] can support our opinion.

4.2. The Extrinsic Method

We evaluate the performances of simple matching similarity, the Ahmad and Dey's similarity and our proposed Hellinger Distance similarity in term of k-modes clustering by comparing the resulting clustering structures to the prespecified structure which is reflects the inherent structure of a classification dataset.

We evaluate our experiments on Vote dataset, Soybean-small dataset, Zoo dataset, Wisconsin Breast Cancer dataset and adult datset. Vote dataset consists of 435 data objects and 16 categorical valued attributes. It has two clusters of Republication and Democrat. Soybean-small dataset consists of 47 data objects and 35 attributes. It has 4 clusters which are labeled as D1, D2, D3 and D4. Zoo dataset consists of 101 data objects an 16 attributes. It has 7 clusters. Wisconsin breast cancer dataset consists of 699 data objects and 9 attributes. It has two clusters of Benign and Malignant. Adult dataset consists of 48842 data objects and 14 attributes. We contain 7 of categorical attributes in experiment. It has 2 clusters.

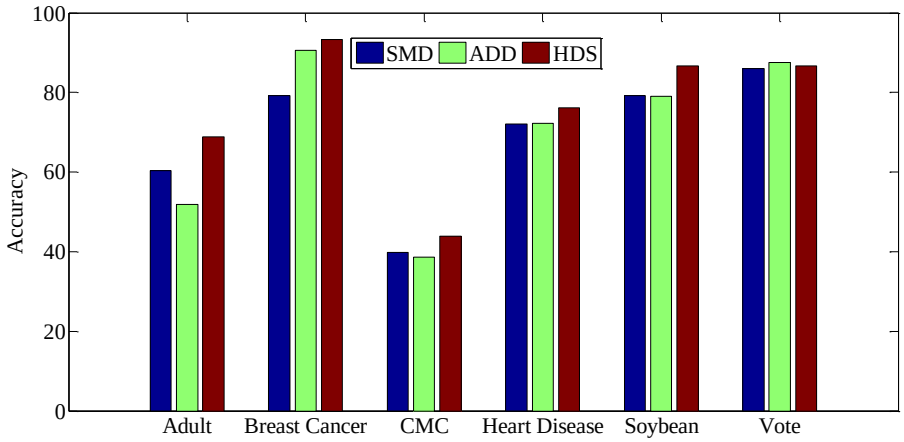

Figure 3. Clustering Comparisons with Acc

Figure 3 reports the results on 6 datasets with different similarity. The evaluations are conducted with SMS, ADD and our proposed similarity. For each dataset the average performance is computed 100 times for every similarity. On Adult dataset which we give an example with in Section 2, ADD's accuracy is lower than the others. It shows the disadvantage of ADD on the datasets which have attributes of unbalance distribution. Vote dataset is actually a binary dataset, all attributes including class label have only 2 values, we think that is why the 3 similarities have nearly the same accuracy on this dataset. Compare with SMS, HDS improve ACC rate range from 0.07% (Vote dataset) to 17.85% (breast cancer), the average ACC improvement is 9.64%. Compare with ADD, HDS improve ACC rate range from -1% (Vote dataset) to 33% (Adult dataset), the average ACC improvement is 10.63%. So we can say that the HDS is better than the ADD and SMS on clustering accuracy.

Table 8. Variance of Clustering Accuracy (10^{-4})

Dataset / similarity	Adult	Breast cancer	CMC	Heart	Soybean-small	Vote
HDS	11	71	2.19	70	277	0.0013
SMS	63	351	8.87	70	243	0.5206
ADD	140	92	18	164	323	0.0133

The second index we use to compare these 3 similarities is variance. In probability theory and statistics, variance measures how far a set of values is spread out. A small variance indicates that the data points tend to be very close to the expected value. In our experiences a small variance means the similarity has a stable clustering result.

Table shows the variance of clustering accuracy. It is easy to find out that variance keep pace with accuracy, most clustering which gets higher accuracy almost takes lower variance. So we can say HDS is also better than ADD and SMS on clustering stability.

5. Conclusion

We propose a new similarity measure based on Hellinger Distance to measure the distance between two categorical variables of an attribute in unsupervised learning. Theoretical analysis and substantial experiments show HDS has better performance with k-modes algorithms not only on accuracy but also on stability. In future, we would like to extend this work to build an unite framework for polymorphic data type and apply on machine learning tasks with mixed dataset.

References

[1] J. Han, M. Kamber and J. Pei, "Data mining: concepts and techniques: concepts and techniques", Elsevier, **(2011)**.
[2] C. Stanfill and D. Waltz, "Toward memory-based reasoning", Communications of the ACM, vol. 29, no. 12, **(1986)**, pp. 1213-1228.
[3] S. Cost and S. Salzberg, "A weighted nearest neighbor algorithm for learning with symbolic features", Machine learning, vol. 10, no. 1, **(1993)**, pp. 57-78.
[4] S. Boriah, V. Chandola and V. Kumar, "Similarity measures for categorical data: A comparative evaluation", red, vol. 30, no. 2, **(2008)**.
[5] G. Gan, C. Ma and J. Wu, "Data clustering: theory, algorithms, and applications", Siam, **(2007)**.
[6] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data", Data & Knowledge Engineering, vol. 63, no. 2, **(2007)**, pp. 503-527.
[7] V. Ganti, J. Gehrke and R. Ramakrishnan, "CACTUS—clustering categorical data using summaries", In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, **(1999)**, pp. 73-83.
[8] Hellinger, E. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. Journal für die reine und angewandte Mathematik, vol. 136, **(1909)**, pp. 210-271.
[9] D. A. Cieslak and N. V. Chawla, "Learning decision trees for unbalanced data", In Machine learning and knowledge discovery in databases. Springer Berlin Heidelberg, **(2008)**, pp. 241-256.
[10] C. R. Rao, "A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance", Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa, vol. 19, no. 1, **(1995)**, pp. 23-63.
[11] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial", Now Publishers Inc., **(2004)**.

[12] I. Csisz, "Information-type measures of difference of probability distributions and indirect observations", Studia Sci. Math. Hunga., vol. 2, **(1967)**, pp. 299-318.
[13] T. Morimoto, "Markov processes and the H-theorem", Journal of the Physical Society of Japan, vol. 18, no. 3, **(1963)**, pp. 328-331.
[14] S. M. Ali, and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another", Journal of the Royal Statistical Society. Series B (Methodological), **(1966)**, pp. 131-142.
[15] D. L. Davies and D. W. Bouldin, "A cluster separation measure", Pattern Analysis and Machine Intelligence, IEEE Transactions on, no. 2, **(1979)**, pp. 224-227.

Authors

Zhao Liang, He received the BS and MS degrees in Computer Science from Liaoning Technical University, China, in 2002 and 2009. He is currently working towards the PhD degree in Liaoning Technical University. His research interests include data mining and machine learning.

Liu Jianhui, He received the BS in Electrical Automation from Fuxin Mining Institute, China, in 1982. He is currently a professor in the Liaoning Technical University. His research interests include computer network and artificial intelligence.