CHEMICAL ENGINEERING TRANSACTIONS
VOL. 46, 2015
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Peiyu Ren, Yancang Li, Huiping Song

ISBN 978-88-95608-37-2; **ISSN** 2283-9216
DOI: 10.3303/CET1546037

Aggregation of Individual Feature Based Similarities and Application to Hierarchical Clustering

Ning Zhou*[a], Bojun Xie[b], Tao Wang[b]

[a] President office,Hebei University, Baoding, 071002,China,
[b] Department of Mathematics and Computer Science,Hebei University,Baoding, 071002,China.
zhouning@hbu.edu.cn

The measure of similarity/dissimilarity is important to clustering algorithms. By using different similarity metrics, a clustering algorithm may achieve different clustering results. Many existed clustering methods employ classic distance measures such as the Euclidean and Manhattan distances as the measures of dissimilarity. In this paper, OWA aggregation operators with learned weighting vectors, such as DOWA and kNN-DOWA, are employed to aggregate the feature-based similarities between instances. The aggregated similarities provide more options in classic clustering algorithms and hence, increase their flexibility. The performances of proposed methods are tested in the classic hierarchical clustering. Experimental results shows that the DOWA and kNN-DOWA aggregated similarities have achieves better clustering accuracies than the Euclidean and Manhattan distances in hierarchical clustering.

1. Introduction

Data mining has become a commonly used technology of data engineering, which is able to solve many real problems (Ren, Liu and Zhang (2015), Wang et al. (2015)). Clustering analysis, in particular, has been widely applied to knowledge discovery, pattern recognition and many applications. The task of clustering is to assign a set of objects into groups (namely clusters) such that the objects in the same group are similar to each other, and dissimilar to those in the other clusters (Jain et al. (1999)). A number of clustering algorithms have been proposed in the literature, with many of them have been successful applied to solve real problems (Jin, Kou and Liu (2014), Górriz et al. (2005).The classic clustering algorithms includes: hierarchical clustering, K-means, fuzzy c-means and so on.

In many clustering algorithms, one of the key parameters is to select the similarity or dissimilarity measurements between instances. Such dissimilarity measurements include the Manhattan distance which calculates the absolute difference of feature (attribute) values of two instances based on each feature, and sum the differences of all the features. One possible drawback of using the classic Manhattan and Euclidean distances as dissimilarity measures in clustering are that all of the features are treated fairly, e.g., has the same weight in the aggregation. Therefore, the reliability of an individual feature is not considered. In real world applications, the reliabilities of different features in a data set can be very different. Many features in data sets contain noise due to inaccurate observation (Nettleton, Orriols-Puig and Fornells (2010)). Therefore, in this paper, the reliability based OWA aggregation operators are employed to aggregate the similarities between instances measured by different features for hierarchical clustering, in order to reduce the interference of noise data and increase the accuracy and robust of clustering algorithms.

The remainder of this paper is organized as follows. Section II introduces the basics of the OWA (Ordered Weighted Averaging), DOWA (Dependent OWA) and kNN-DOWA (k-Nearest Neighbor DOWA) aggregation operators. Section III defines the reliability based aggregation of similarities observed from individual features and describes its application to hierarchical clustering. Section IV presents the experimental evaluation of the proposed approach and discusses the results. Finally, Section V concludes the paper and points out directions for further development.

Please cite this article as: Zhou N., Xie B.J., Wang T., 2015, Aggregation of individual feature based similarities and application to hierarchical clustering, Chemical Engineering Transactions, 46, 217-222 DOI:10.3303/CET1546037

2. Background

Aggregation of several input values into a single output value is an indispensable tool not only for mathematics or physics, but for many real-world applications in engineering, economic, social, and other fields. It is also a useful tool in many real applications of information sciences such as group decision making (Xu (2006)), human resource management (Canós and Liern (2008)), and journal ranking (Su (2014)). Apart from the classical aggregation operators (such as average, maximum and minimum), another interesting and more general type of aggregation operator is the Ordered Weighted Averaging (OWA) operator. OWA is a family of aggregation operators which are parameterized based on the ordering of the inputs. The fundamental aspect of this family of operators is the reordering step in which the inputs are rearranged in descending order and then integrated into a single aggregated value.

Definition 2.1 (Yager(1988)): A mapping OWA: $\mathbb{R}^n \to \mathbb{R}$ ($\mathbb{R}$ is the set of real number) is called an OWA operator if

$$\mathrm{OWA}(a_1, a_2, \cdots, a_i, \cdots, a_n) = \sum_{j=1}^{n} w_j b_j \tag{2.1}$$

Where b_j is the j-th largest number in $(a_1, a_2, \cdots, a_i, \cdots, a_n)$, $W = (w_1, w_2, \cdots, w_n)$ satisfies that $w_j \in [0,1]$ for each $j \in \{1, 2, \cdots, n\}$, and $\sum_{j=1}^{n} w_j = 1$. $(a_1, a_2, \cdots, a_n)$ is called the input vector, W is the weighting vector.

One feature of the OWA operator is: the input values in $(a_1, a_2, \cdots, a_i, \cdots, a_n)$ are descending ordered and form a new vector $(b_1, b_2, \cdots, b_j, \cdots, b_n)$ before aggregation with W. In other words, the weights are assigned to the rearranged inputs, and then weighted-summed. Compared with the simple weighted-sum aggregation, the weight w_j is not directly assigned with the original input a_i, but assigned to the j-th position in the reordered input vectors. A common example of the OWA operator in real application is: remove the maximum and minimum inputs and average the remains, which is according to an OWA operator with the weighting vector:

$$W = (0, \frac{1}{n-2}, \frac{1}{n-2}, \cdots, \frac{1}{n-2}, \frac{1}{n-2}, 0). \tag{2.2}$$

Besides that, the commonly used special cases of OWA operator include the maximum, minimum and average.

Although such an OWA-based weighted aggregation is straightforward and has been widely accepted, it under-weighted the role of the maximum and the minimum in decision-making, while eliminated the difference of importance among other input values (Wang and Xu (2008)). Therefore, Xu (2005) and Boongoen and Shen (2010) proposed two OWA operators which learn weights of the inputs from the input values based on their reliabilities. The two methods are named as the DOWA and kNN-DOWA operators, respectively.

Definition 2.2 (Xu (2005)): Let μ be the mean of the inputs, $s(a_i, \mu) = 1 - (|a_i - \mu| / \sum_{j=1}^{n} |a_j - \mu|)$ be the similarity between an input value to μ. A mapping DOWA: $\mathbb{R}^n \to \mathbb{R}$ is called a DOWA operator if

$$\mathrm{DOWA}(a_1, a_2, \cdots, a_i, \cdots, a_n) = \sum_{i=1}^{n} w_i a_i \tag{2.3}$$

where $w_i = s(a_i, \mu) / \sum_{j=1}^{n} s(a_j, \mu), i \in \{1, 2, \cdots, n\}$ is the weight learned from the input vector.

Definition 2.3 (Boongoen and Shen (2010)): Let $\{n_1^i, \cdots, n_t^i, \cdots, n_k^i\}$ be the k nearest neighbors of the input value a_i found in the input vector, and the reliability of a_i is $R^k(a_i) = 1 - \sum_{t=1}^{k} |a_i - n_t^i| / kD_{\max}$, ($D_{\max}$ is the maximum distance between all the input values), then a mapping kNN-DOWA: $\mathbb{R}^n \to \mathbb{R}$ is called a kNN-DOWA operator if:

$$\text{kNN-DOWA}(a_1, a_2, \cdots, a_i, \cdots, a_n) = \sum_{i=1}^{n} w_i^k a_i \tag{2.4}$$

where $w_i^k = R^k(a_i) / \sum_{j=1}^{n} R^k(a_j)$, $i \in \{1, 2, \cdots n\}$ is the weight learned from the input vector.

3. Aggregation of feature based similarities

In general, a data set used in pattern recognition and machine learning contains many features (attributes). These attributes can be used to discover the degree of similarity or dissimilarity between the instances in the data set. Dissimilarity degree is normally expressed by using the distance between instances, while the degree of similarity can be expressed as a decreasing function of dissimilarity (such as reciprocal), which is usually normalized to [0, 1]. Since each instance can have different attribute values in different features, a pair of instances can have different degrees of similarity evaluated by different features. As it is mentioned above, not all of the features are reliable due to the existence of noise data. Therefore, this paper employs DOWA and kNN-DOWA to calculate the similarity metric in hierarchical clustering algorithm in order to eliminate the effect of noise data. For each pair of instances in a data set, the similarities between them are estimated based on each individual feature based on their values on that feature. The DOWA or kNN-DOWA operator is applied to aggregate the similarities evaluated on different feature. The aggregated results can be deemed as an all overall similarity degree between the pair of instances and are applied to clustering algorithms.

3.1 Aggregation of Similarities

An example extracted from the Iris data set Fisher (1936)) is employed to demonstrate the aggregation of similarities based on individual feature. The Iris data set is perhaps the best known database to be found in the pattern recognition literature. It is a classic in the field and is referenced frequently to this day. The Iris data set have four conditional features, which are A1: sepal's length, A2: sepal's width, A3: petal length and A4: petals width. It also includes a feature which indicated the class label of each instance. Table 1 shows a subset of the Iris data set which contains three instances.

Table 1: Three Examples from Iris Data Set

No.	A_1	A_2	A_3	A_4	Class
x_1	5.6	3.0	4.1	1.3	Iris-versicolor
x_2	6.6	3.0	4.4	1.4	Iris-versicolor
x_3	6.0	2.2	5.0	1.5	Iris-virginica

Using the absolute difference of attribute values as an example of dissimilarity metric on each individual feature, it can be calculated that the degrees of dissimilarity between x_1 and x_2 on the four given features are $d_{A_1 \cdots A_4}(x_1, x_2) = (1.0, 0.0, 0.3, 0.1)$. It can be seen from the example that when evaluating the dissimilarity between x_1 and x_2, the value given by feature A1 is significantly greater than the values given by the other three features. Using the known class label as a reference, one can confirm that x_1 and x_2 belong to the same class and hence, their degree of dissimilarity should be relatively small. Therefore, the degree of dissimilarity given by feature A1 not only conflicts with those given by other features, but also conflicts with the class labels. Similarly, when evaluating the degree of dissimilarity between x_2 and x_3, $d_{A_1 \cdots A_4}(x_2, x_3) = (0.6, 0.8, 0.6, 0.1)$ the value given by feature A4 is much smaller than values given by the other three features, and also conflicts with the class label. Therefore, the DOWA and kNN-DOWA operators are employed in this paper to aggregate the individual feature based dissimilarities between instances, e.g.:

$$d_{\mathrm{DOWA}}(x_a, x_b) = \mathrm{DOWA}(d_{A_1}(x_a, x_b), \cdots, d_{A_n}(x_a, x_b)) \tag{3.1}$$

$$d_{\mathrm{kNN}}(x_a, x_b) = \text{kNN-DOWA}(d_{A_1}(x_a, x_b), \cdots, d_{A_n}(x_a, x_b)) \tag{3.2}$$

where x_a and x_b are two different instances in a data set. Using the examples in Table 1 as well, when the DOWA operator is applied, the weighting vectors which with respect to $d_{A_1 \cdots A_4}(x_1, x_2)$ and $d_{A_1 \cdots A_4}(x_2, x_3)$ are (0.1667, 0.2436, 0.3205, 0.2692) and (0.3039, 0.2255, 0.3039, 0.1667), respectively. In a similarly way,

when kNN-DOWA (k=1) is applied, the corresponding two weighting vectors are (0.188, 0.275, 0.261, 0.275) and (0.286, 0.250, 0.286, 0.184), respectively. By using the operators which consider the reliability of inputs, the "noise" values such as $d_{A_1}(x_1, x_2)$ and $d_{A_4}(x_2, x_3)$ can be weighted less than other normal values in the aggregation of feature based similarities.

3.2 Applying the Aggregated Similarity to Hierarchical Clustering

In addition to using different operators to aggregate the different similarities estimated on different individual features, the methods used to measure the dissimilarity between two instances on an individual feature can also be different. It is worth noticing that this paper only examines the influence of different aggregation operators in constructing the overall similarity between instances, thus, the evaluation of each individual feature based similarity simply employs the absolute difference between the attribute values: $d_{A_i}(x_a, x_b) = | A_i(x_a) - A_i(x_b) |$, where $A_i(x_a)$ is the attribute value of x_a in attribute A_i. It is worth noticing that other similarity metrics can also be employed.

Hierarchical clustering is one of the most significant developments in clustering algorithms. In particular, hierarchical clustering builds a cluster hierarchy or a tree/dendrogram of clusters. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity.

The main reason for using the hierarchical clustering to test the proposed aggregated fuzzy relations is that any forms of similarity or distance can be applied to the hierarchical clustering directly. Consequently, the clustering results are mainly dependent on the weights employed in the proposed aggregations of similarities. Given a data set with *N* instances, *n* attributes, the hierarchical clustering with DOWA or kNN-DOWA aggregation operator can be fulfilled in the following steps:

Step1. Using Equation (3.1) or (3.2) to calculate the dissimilarity between two instances, considering the symmetric and reflexive properties, there are *N*(*N-1*)*/2* dissimilarity values to be calculated;

Step 2. Initialize each instance to a cluster, and *N* clusters will be gained, each of which contains only one instance;

Step 3. Find the pair of clusters which share the link with the smallest value and then merge them to form a new cluster;

Step 4. Update the values of links between the new cluster and all the old clusters;

Step 5. Repeat step 3 and 4, till there is only one cluster (or *m* clusters, if *m* is given);

where the link (in Steps 3 and 4) defines the way how the hierarchical clustering algorithm characterizes the similarity between a pair of clusters. The popular options of link in hierarchical clustering algorithms are the single-link and complete-link. In single-link, the link between two clusters is the minimum of the dissimilarity between two instances drawn from the two clusters, respectively. In complete-link, the link between two clusters is the maximum of all pairwise dissimilarity between instances in the two clusters. In either case, two clusters are merged to form a larger cluster based on minimum link criteria. The complete-link algorithm produces tightly bound or compact clusters while the algorithm suffers from a chaining effect Jain (1999)). Due to these reasons, the complete-link hierarchical clustering is used in the following experiment.

4. Experiments and results

In order to examine the performance of DOWA and kNN-DOWA operators in hierarchy clustering, the conventional Euclidean and Manhattan distances based hierarchical clustering algorithms are employed to conduct comparison. When calculating the reliability of inputs in kNN-DOWA, the number of neighbors is set to ceil of *n*/2. All these tested data sets contain only numeric conditional features, and each conditional feature in each data set in normalized to [0, 1]. The tested data sets are drawn from the UCI Machine Learning Repository (Frank and Asuncion (2010)). The results of clustering are evaluated by accuracy, where the known class labels of instances are employed as ground truth. Each data set is grouped into *m* clusters, where *m* is set to the number of known classes in each data set. The accuracy of clustering results are shown in Table 2, where each number is based on only one time experiment, since the single-link and complete-link hierarchical clustering do not contain random parameters.

The experimental results show that: By using the DOWA and kNN-DOWA operators, the accuracy of hierarchical clustering is better than that of Euclidean or Manhattan distance based clustering. In three of the five tested data sets, DOWA operator achieved best accuracy. For kNN-DOWA operator, the parameter *k* is fixed to *n*/2 in this experiment, which may affect the result of kNN-DOWA. Therefore, Table 3 lists the clustering accuracies achieved by kNN-DOWA when *k* is set to 1/4*n*, 2/4*n*, 3/4*n* and 4/4*n*. It can be seen from Table 3 that kNN-DOWA has the potential of achieving higher accuracy, if *k* is properly selected. However, the change of accuracy is not monotonic with the change of *k*. Therefore, how to choose the best value /range of *k* for this application is potential direction to work on.

Table 2: Accuracy with Complete Link

Data set	Euclidean	Manhattan	DOWA	kNN-DOWA
Iris	88.00%	84.00%	82.67%	84.67%
Wine	93.26%	94.38%	96.63%	94.38%
Glass	51.87%	45.33%	52.34%	56.54%
Ecoli	75.89%	77.38%	78.87%	76.79%
Heart	56.30%	73.33%	75.56%	70.00%

Table 3 kNN-DOWA Accuracy with Change of k

Data set	k=1/4n	k =2/4n	k =3/4n	k =4/4n
Iris	77.33%	84.67%	88.67%	-
Wine	95.51%	94.38%	93.82%	97.75%
Glass	56.07%	56.54%	50.00%	50.00%
Ecoli	77.38%	76.79%	76.49%	76.79%
Heart	57.41%	70.00%	69.26%	76.67%

5. Conclusions

In this paper, OWA aggregation operators with learned weighting vectors, i.e., DOWA and kNN-DOWA are employed to aggregate the feature-based similarities between instances. Through the introduction, analysis on a famous data example (Fisher's Iris) and experimental results, it can be seen that by using the DOWA and kNN-DOWA operators, the hierarchical clustering results are better than those of using classical distances in term of accuracy.

However, the learning of reliability-based weights in aggregation operators needs extra computation. Further improvements in time efficiency of these methods are desirable. Besides, the best selection of k in the kNN-DOWA is also a challenging problem.

References

Boongoen T. and Shen Q., 2010, Nearest-neighbor guided evaluation of data reliability and its applications. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 40(6): 1622-1633 DOI: 10.1109/TSMCB.2010.2043357

Canós L. and Liern V., 2008, Soft computing-based aggregation methods for human resource management. European Journal of Operational Research 189(3): 669–681 DOI: 10.1016/j.ejor.2006.01.054

Fisher R.A., 1936, The use of multiple measurements in taxonomic problems. Annals of Eugenics 7(2): 179–188 DOI: 10.1111/j.1469-1809.1936.tb02137.x

Frank A. and Asuncion A., 2010, UCI machine learning repository. Available in URL: http://archive.ics.uci.edu/ml/datasets.html

Górriz J.M., Ramírez J., Lang E.W. and Puntonet C.G., 2005, Hard C-means clustering for voice activity detection. Speech Commun 48(12):1638–1649 DOI: 10.1016/j.specom.2006.07.006

Jain A.K., Murty M.N. and Flynn P.J., 1999, Data clustering: a review. ACM Comput Surv 31(3):264–423 DOI: 10.1145/331499.331504

Jin R., Kou C. and Liu R., 2014, Biclustering algorithm of differential co-expression for gene data. Review of Computer Engineering Studies 1(1): 7-14 DOI: 10.18280/rces.010102

Nettleton D.F., Orriols-Puig A. and Fornells A., 2010, A study of the effect of different types of noise on the precision of supervised learning techniques, Artificial Intelligence Review 33(4): 275–306 DOI: 10.1007/s10462-010-9156-z

Ren X., Liu Q. and Zhang Y., 2015, The proportion of energy consumption structure prediction based on markov chain. Mathematical Modelling and Engineering Problems 2(1): 1-4 DOI: 10.18280/mmep.020101

Su P., Chen T., Shang C., and Shen Q., 2014, Nearest neighbour-guided induced OWA and its application to journal ranking. 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE): 1794-1800 DOI: 10.1109/FUZZ-IEEE.2014.6891805

Wang L., 2014, An ungreedy Chinese deterministic dependency parser considering long-distance dependency. Review of Computer Engineering Studies 1(2): 1-4 DOI: 10.18280/rces.010201

Wang T., Yan H., Zhong S. and Zhang Y., 2015, Research of fire alarm system based on extension neural network. Review of Computer Engineering Studies 2(1): 9-14 DOI: 10.18280/rces.020102

Wang Y. and Xu Z., 2008, A new method of giving OWA weights. Mathematics in Practice and Theory 38(3): 51-61

Xu Z., 2006, Induced uncertain linguistic owa operators applied to group decision making. Information Fusion 7(2): 231–238 DOI: 10.1016/j.inffus.2004.06.005

Xu Z.S., 2005, An overview of methods for determining OWA weights. International Journal of Intelligent Systems 20(8): 843-865 DOI: 10.1002/int.20097

Yager R.R., 1988, on ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Transactions on Systems, Man and Cybernetics 18(1): 183-190 DOI: 10.1109/21.87068