VARIABLE-SCALE CLUSTERING FOR DECISION MAKING

Xuedong Gao, Ai Wang*
Donlinks School of Economics and Management
University of Science and Technology Beijing
Beijing, China
wangai22222@126.com

ABSTRACT

Decision making can be regarded as a problem-solving activity and decision makers usually consider a problem from different perspectives, hierarchies, dimensions, that is referred to as scale transformation (ST). Although the research on ST related to clustering have achieved some progress, it is mainly confined to the geography and image area. This paper mainly studies the ST problem among clustering analysis especially for decision making. We establish the scale transformation rate (STR) to measure the effect of ST based on the rough set theory. What's more, a variable-scale clustering algorithm (VSC) is also proposed. Experiment illustrates that comparing to the k-means, the VSC is able to ensure every cluster are qualified (which is not limited to only optimize the overall performance) and shows great potential in decision making.

Keywords: variable-scale clustering, concept hierarchy, rough set theory.

1. Introduction

Clustering analysis is one of the traditional unsupervised learning methods, which aims to group a set of objects without predefined classes (Wu, S., Gao, X.D., & Bastian M., 2003). Formally, let X denote the instance space, the task of clustering analysis is to partition X into several clusters C_i, where (1) $C_i \neq \emptyset$, (2) $\cup C_i = X$, (3) $C_i \cap C_j = \emptyset$. Here, $x_i \in X$ is an instance characterizing the attributes (features) of an object and x_i in the same cluster C_i are more similar to each other than to those in other clusters.

Clustering has developed into a powerful data analysis tool and is widely used in many research fields (e.g. customer analysis, pattern recognition), especially decision making. Decision making can be regarded as a problem-solving activity and its task is to rank different alternatives in terms of how attractive they are to the decision-makers when all the evaluative criteria are considered. It can be seen that the decision-making process considers a problem from different perspectives, hierarchies, dimensions, that is referred to as scale transformation (ST). Thus, clustering should support this ST demand. Although the research on scale transformation related to clustering have achieved some progress, it is mainly confined to the geography and image area.

The ST problem of clustering specially for decision making suffers from these challenges: (1) lack of the standard definition of scale. The scale among traditional vector space is not as obvious as the time or space attribute in the geography and image field. (2) there is no appropriate metric to quantitatively evaluate the effect of ST. (3) variable-scale mechanism according to the thinking process of decision makers should be established to improve the efficiency of ST.

This paper focuses on the scale transformation problem among clustering analysis. The main contributions are as follows. First, we illustrate the definition of scale among clustering analysis based on the concept hierarchy theory, which is consistent with the multi-scale clustering domain. Second, we propose the scale transformation rate (STR) to measure the effect of ST (i.e., scaling-up or scaling-down concept) based on the rough set theory. Third, a variable-scale clustering algorithm (VSC) is proposed. A case study on student training-program design is conducted to verify the effectiveness of the VSC. Experiment results show that the VSC is able to ensure every cluster are qualified (satisfied) for decision makers in contrast to the classic clustering method k-means (which only concerns the overall optimization objective).

2. Literature Review

Multi-scale clustering is a popular area, which aims to transform the clustering results (knowledge) on the current scale directly to other scale based on the multiscale technique. However, the definition of scale has not yet formed a complete theory or method.

Sun, D.H. (2015) considers that the scale is an equivalence relation describing a concrete dataset. From the perspective of information granule, the scale reflects the information carried by the research object. Large scale is more likely to describe the macro feature, while small scale describes the microscopic feature. Han, Y.H. (2016) considers that the scale is a unit of concept range describing a concrete object. These concepts follow the partial order relation among the corresponding concept hierarchy. Combining the views above, we obtain the definition of scale in the ST problem.

Definition 1. Given a dataset D, a variable-scale dataset $D^S = (U, OA, TA, V, f)$, where U is the universe of D, OA is the original attribute set (original scale) of D, V_a is the value of attribute $a(a \in OA)$, where information function $f: U \times OA \to V$, $V = \cup V_a$, and TA is the target attribute set (target scale) of D, V_t is the value of attribute $t(t \in TA)$, where $\forall TA$, $TA \preccurlyeq OA \to V_t \preccurlyeq V_a$ ($\forall TA, TA \succcurlyeq OA \to V_t \succcurlyeq V_a$), $V_t \notin V$.

According to the Def.1, ST is a process of transforming the original scale OA to the target scale TA on the same dataset D following the partial order relation of a known concept hierarchy.

3. Methodology

The rough set theory is a mathematical tool to deal with uncertain knowledge and has provided many metrics to measure the classification ability of an attribute (set). Consequently, we define the scale transformation rate (STR) to measure the effect of ST based on the rough set theory.

Definition 2. The scale transformation rate (STR) is:

$$\mathrm{STR}(OA, TA) = \sum_{i=1}^{n} |TA\_(OA_i)| \,/|U| \qquad (1)$$

$$TA\_(OA_i) = \cup \{TA_j | TA_j \subseteq OA_i\} \qquad (2)$$

Where $U/OA = \{OA_1, OA_2, \cdots, OA_n\}$, $U/TA = \{TA_1, TA_2, \cdots, TA_m\}$.

Note that $\mathrm{STR}(OA, TA)$ reflects the approximate quality of TA to OA, $\mathrm{STR}(OA, TA) \in [0,1]$. If $\mathrm{STR}(OA, TA) = 1$, it means $\forall TA_j \to (\exists OA_i)\ TA_j \subseteq OA_i$.

There are various metrics (e.g. RMSSTD, Dunn, Hubert) to evaluate the results of a clustering method. Since the task of variable-scale clustering is to obtain the partition results where all the clusters satisfy decision makers, we select the RMSE to evaluate the quality of each cluster (see Eq. (3)).

$$\mathrm{RMSE}(C_i) = \sqrt{\sum_{x_j \in c_i} dist(x_j, c_i)^2 / |C_i|} \tag{3}$$

Where c_i is the centroids of cluster C_i, $dist(x_j, c_i)$ is the distance between the instance x_j and c_i, and we apply the Euclidean distance to calculate the similarity between different instances in Section 4.

We propose a variable-scale clustering algorithm (VSC) and the pseudo code is shown in Algorithm 1. The time complexity of VSC is $O(nkt)$, where n is the number of instances, k is the number of clusters, and t is the number of iterations.

Algorithm 1 *VariableScaleClustering* (*D*, S_0,*k*)

```
C=D.initialCluster (k)
D.delete ($C_i.qualified$)
for $D \neq \emptyset$ do
    for all $A_j \in D$ do
        if $\mathrm{STR}(A_j, TA) < S_0$ then
            D.update ($A_j, TA$)
        break
        end if
    end for
    $R_0 = RMSE(C_i.closestQualified)$
    C=D.Cluster (k-count($C_i.qualified$))
    for all $C_i \in C$ do
        if $RMSE(C_i) < R_0$ then
            D.delete ($C_i$)
        end if
    end for
end for
```

4. Experiment Results and Discussion

A case study has been provided here to illustrate the performance of the proposed method VSC. A competition committee plan to organize a train-program for participants and invite (about six) experts to give lectures for different participant groups. Hence, the problem is how to divide participants into several clusters with similar and accessible feature, that makes committee easier to decide which professor should be invited.

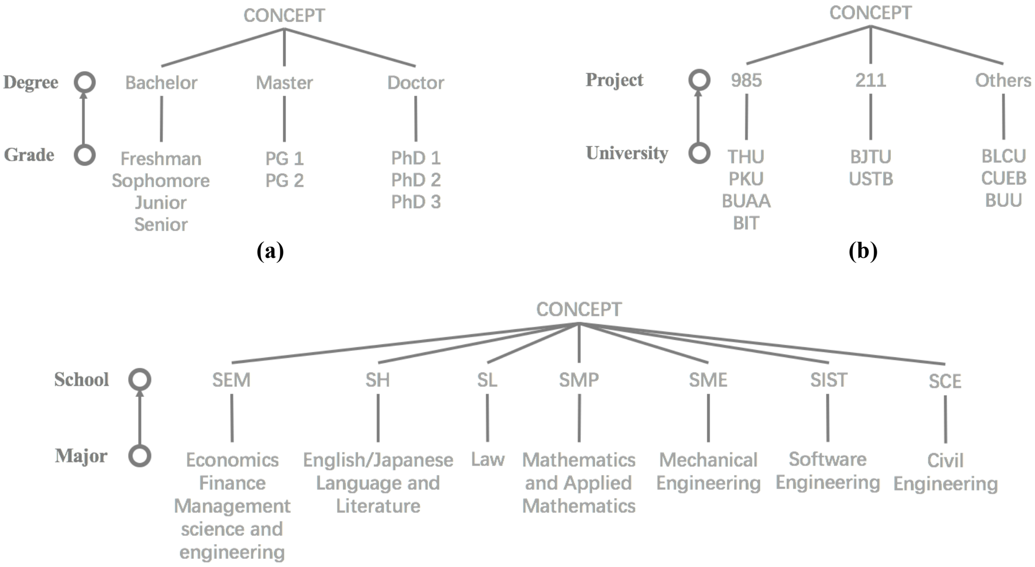
We chose sixty-three samples from the real participant dataset (see Table 1) and the concept hierarchy of every attribute (see Fig.1) were established through the prior knowledge of researchers (which has been accepted by committee members). Moreover, the measure RMSE was applied to evaluate the accuracy of each cluster by Eq. (3). Experiments were performed in OS X (10.11.3) environment on a machine with 8GB RAM. All methods were coded in Weka (3.8.1).

Fig.2 shows the results of the VSC and k-means. The color reflects the value of RMSE, green represents qualified clusters, yellow represents the closest-qualified clusters (boundary of a certain scale), while red represents unqualified clusters for decision makers. Consequently, the value of RMSE is steadily increasing from deep green to deep red. What's more, the width of rectangle represents the number of instances in each cluster.

Table 1
Data description.

Attribute	Value
University	Tsinghua University (THU), Peking University (PKU), Beihang University (BUAA), Beijing Institute of Technology (BIT), Beijing Jiaotong University (BJTU), University of Science and Technology Beijing (USTB), Beijing Language and Culture University (BLCU), Capital University of Economics and Business (CUEB), Beijing Union University (BUU)
Major	Economics, Finance, Management Science and Engineering, English Language and Literature, Japanese Language and Literature, Law, Mathematics and Applied Mathematics, Mechanical Engineering, Software Engineering, Civil Engineering
Grade	Fresh man, Sophomore, Junior, Senior, postgraduate student in the first year (PG 1), postgraduate student in the second year (PG 3), Ph.D. student in the first year (PhD 1), Ph.D. student in the second year (PhD 2), Ph.D. student in the third year (PG 3),

According to Table 1 and Fig.1, there are three attributes and each attribute exists two scales, which forms eight (2^3=8) possible ST approaches. At the beginning, the VSC discovered two qualified (satisfied) clusters on the original scale (i.e., University, Major, Grade) confirmed by committee members. After that, the VSC identified the closest-qualified cluster and improved the concept *Major* to the higher concept *School*. We got the boundary (threshold) of the next clustering through calculating the RMSE of the closest-qualified cluster on this target scale. It can be seen that the closest-qualified cluster has reached the requirement at the next clustering with the help of scale improvement, meanwhile, its actual meaning is able to be accurately described by the target scale. In the same way, the VSC finally divided the sixty-three participants into seven clusters corresponding to three scales. At this point, the committee are able to clearly decide which professor is more appropriate for this training-program through each cluster and its scale feature.



(a)

(b)

(c)

Fig.1. Concept hierarchy of the participant dataset.

Since the committee planned to invite about six experts before and the VSC finally obtained seven qualified clusters, we compared the results to the classic clustering method k-means with the initial parameter k=6 or 7. Take the maximum RMSE of qualified clusters on each scale as the threshold, k-means always got unqualified clusters although the number of unqualified clusters decreased as the increasing of scales and initial parameter k. Thus, the VSC is more accurate and stable for practical applications, especially for decision making.

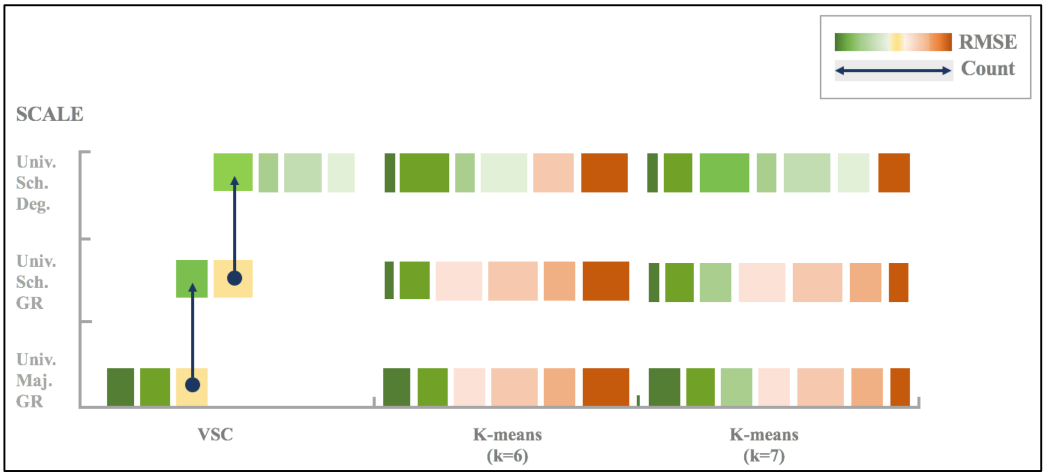


Fig. 2. The comparison results of the VSC and k-means.

5. Conclusions

Variable-scale clustering is an emerging research field and mainly studies the scale transformation (ST) problem among the clustering analysis. First, this paper defined the core concept *Scale* in the ST problem based on the concept hierarchy theory, since there is no standard definition in previous research. Second, this paper proposed the scale transformation rate (STR) on the basis of the rough set theory, which solves the problem of the quantitative evaluation on ST. Third, a variable-scale clustering algorithm (VSC) is proposed. Experiment illustrates that comparing to the k-means, the VSC is able to ensure every cluster are qualified (which is not limited to only optimize the overall performance) and shows great potential in decision making.

However, the VSC also has the initial value selection problem on the parameter (i.e., scale transformation threshold S_0). The future work will focus on designing the variable-scale mechanism to provide effective methods for threshold setting.

6. Key References

Wu, S., Gao, X.D., & Bastian M. (2003). *Data Warehouse and Data Mining*. Beijing: Metallurgical Industry Press.

Sun, D.H., & Zhao, S.L. (2015). Weight vector based multi-scale clustering algorithm. *Computer Science, 42*, 263–267.

Han, Y.H., & Zhao, S.L. (2016). Multi-scale clustering algorithm. *Computer Science, 43*, 244–248.