Information-Theoretic Based Clustering Method for High-Dimensional Data

Xuan Huang¹, Lixi Chen²,* and Yinsong Ye³

¹School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China
²School of Electrical Engineering, Southwest Jiaotong University, Chengdu, China
³School of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China

*Corresponding author e-mail: lixi.chen@my.swjtu.edu.cn

Abstract. Clustering is one of the important techniques for pattern recognition. Traditional clustering methods often group data objects by distance measurement. However, it is difficult to measure data in high-dimensional space. This paper proposes a clustering method for high-dimensional data. It combines the information theory criteria to establish clustering rules. The improved of K-Means is used to generate basis clustering, then the clustering ensemble is used to integrate the initial clusters to obtain a more stable final result. Our proposed method is tested on standard datasets, and its performance is compared with K-means ensemble. The experimental results indicate effectiveness for high-dimensional data clustering.

Keyword: Feature Selection, Clustering Ensemble, K-Means, Information-Theoretic

1. Introduction

In the past decades, with the advancement of collection technology, data collection has become easier and the scale of the collected data has continued to expand, and a large number of high-dimensional data have appeared in many fields. Because high-dimensional data is ubiquitous, its research is of great significance.

Clustering is an unsupervised machine learning method. It divides the data set according to its characteristics, making data objects grouped into the same cluster more similar to data
objects in different clusters. In the initial stage of data analysis, clustering technique allows us to discover the unknown world, which has attracted the attention of scientists. It is one of the critical technologies of artificial intelligence and pattern recognition, and has wild applications in many fields, such as image segmentation and file recovery, etc. The sparse of high-dimensional data and the difference between data objects make traditional clustering algorithms that are measured by distance or similarity often invalid, which forms “curse of dimensionality”. Recently, clustering for high-dimensional data have received more attention.

At present, there are mainly two types of clustering research for high-dimensional data. One is to transform the original data feature set into a small number of essential features through dimensionality reduction technology [11], and then use some similarity measurement methods, such as Euclidean distance, to measure data instances for clustering. These algorithm are based on the minimum variance criterion. Another method is subspace clustering, which is a type of clustering method for data in the whole space, which is used to identify the possible clusters in the data set and the subspaces in which these clusters are located. For more detailed subspace clustering techniques, please refer to the reference [7].

For high-dimensional data, the underlying data structure is unknown, and it is difficult to measure the similarity between objects using distance. There are multiple concepts in information theory that can be used to measure the correlation between the objects, and to measure the distance between distributions. Information theoretic measure of similarity can capture data structures other than variance statistics, it can rescue in the important situations where non-negative co-occurrence data is available [4]. This paper proposes an information-theoretic based clustering method for high-dimensional data.

The remaining of this paper are described into five parts. Section 2 introduces the related work such as cluster integration and mutual information, etc. In Section 3 we introduces the concepts of information theory used in this article in detail. In Section 4, presents the information-theoretic based clustering algorithm for high dimensional data. Our experimental details and comparisons are given in Section 5. Section 6 summarizes the paper.

2. Related work
Since the data structure is unknown, existing clustering algorithms try to divide the data with different similarity measures methods. Information theoretic measure of similarity can capture data structures other than variance statistics, it can rescue in the important situations where non-negative co-occurrence data is available [4]. Information entropy is a rather abstract concept representing the uncertainty probability of discrete random events. It is often used to measure the uncertainty of a single variables. Entropy can represent the value of information, it can be used as a measure of the information value of variables, that is, the larger the value of entropy, the more information the variable contains. Mutual information (MI) is also a popular used metric in information theory. It can measure the interdependence between two random variables or sets. These information theory correlation measures can not only measure the importance of features, but also use it to measure the similarity between two different cluster partitions. More detailed definitions of related concepts in section 3.

Clustering is an effective technique for discovering the unknown world in the early stages of data analysis. However, there is no clustering method can effectively applied to all different types of data sets. Even setting different parameters in the same model will get different clustering results. These factors greatly limit the analysis of real world data.
Clustering ensemble analyzes the data set from different perspectives, and consensus multiple base clustering to obtain a single result with more stable and better generalization performance [1]. Especially for high-dimensional data, you can reduce the complexity of the algorithm by grouping the data into different subspaces, clustering in the subspace and then using the consensus function to integrate. This parallel clustering method helps reduce the complexity of the algorithm. Therefore, for high-dimensional data, clusters integration is of great significance. More detailed methods and applications can refer to Ref. [2] [9].

The K-Means’ idea is relatively simple: for a given set of dataset, it is divided into K clusters based on the distance from the objects to the centroid, so the instances in the same cluster are more similar than the objects in different clusters [3][6]. The K-Means has high convergence speed, and good clustering effect, and is widely used. Data are sparse in high-dimensional space, and it is difficult to directly measure their distance. Therefore, K-Means is not suitable for high-dimensional data. Dhillon propose a new information-theoretic divisive algorithm, which is similar to the K-means algorithm, except that it uses KL divergence to measure distance instead of Euclidean distance [6].

3. Some Concepts from Information Theoretic

In this section, we will review some information theory concepts that will be used heavily in this article. The first is entropy, which is used in information theory to measure the value of random variable information. Suppose discrete variable $X$ obeys probability distribution $p(x)$, where $p(x)$ is the probability mass function of $X$. The entropy of $X$ can be defined as [6]:

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

(1)

Where $X$ is represented as a set containing all $x$ elements. Another important concept is Kullback-Leibler (KL) divergence [], which is a measure of the asymmetry of two probability distributions. Suppose $f(x)$ and $g(x)$ are probability density functions of $x$, the KL divergence is defined as:

$$KL(f || g) = \sum_{x \in X} f(x) \log \frac{f(x)}{g(x)}$$

(2)

When we use one distribution to approximate another, it calculates the amount of information lost. KL divergence is a commonly used measure, but it is difficult to estimate it directly in a nonparametric manner in practical applications. Erhan et al combined the properties of multi-dimensional Gaussian functions under Renyi’s entropy, assuming data points $D = \{d_i\}_{i=1}^n$ follow multidimensional Gaussian distribution, the Gaussian kernel denote as Gaussian kernel for probability density function, and substitute it into the formula to get Renyi’s quadratic entropy as follow, Detailed derivation process can refer to Ref. [4].

$$H^\gamma(P) = (2^{\gamma+1} - 1)\gamma\left(\sum_{i=1}^n \rho_i^\gamma - 1\right)$$

(3)
Where \( s > 0, \ s \neq 1 \)

When \( S = 2 \), Eq. (4) is called quadratic entropy, it can be accurately calculated directly from the sample instead of integration. They also derive general mutual information [4], defined as follows:

\[
I^s(x; y) = H^s(y) - H^s(y \mid x)
\]

Mutual Information is a useful measure of information in information theory. It can be seen as the amount of information contained in one random variable about another random variable. For two variables \( X \) and \( Y \), their joint distribution is \( p(x, y) \) and the edge distribution are \( p(x) \) and \( p(y) \), the mutual information between \( X \) and \( Y \) can be defined as:

\[
I(X; Y) = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)}
\]

(5)

It can be seen that if \( X \) and \( Y \) are random variables that are independent of each other, then their mutual information equals to 0.

4. Entropy based clustering method

In the following, we will introduce the following uniform notation of the ensembles clustering problems. Let \( D = \{d_i\}_{i=1}^n \) be a set of \( n \) data objects, where each \( d_i = (d_{i1}, d_{i2}, \ldots, d_{in}) \in R^n \). Also, suppose \( P = \{P_1, P_2, \ldots, P_p\} \) is a set of base partitions of \( D \), each component \( P_j = \{C_{j1}, C_{j2}, \ldots, C_{jk}\} \) denotes the \( j \)-th group, which is consists of a set of nonempty and disjointed cluster, and \( r \) is the index of clusters in \( P_j \), i.e. \( C_{jz} \) (\( z = 1, \ldots, r \)) denotes the \( z \)-th cluster of the \( j \)-th partition. \( \forall P_j \in P, \cup_{j=1}^k C_{jz} = D \), and each component partition in \( P_j \) is disjoint, i.e. \( \forall C_{ja}, C_{kb} \in P_j, s.t. a \neq b, C_{ja} \cap C_{kb} = \phi \).

| notation    | annotation                  |
|-------------|----------------------------|
| \( D \)     | Data set                    |
| \( D' \)    | Data set after dimensionality reduction |
| \( F_0 \)   | Original feature set        |
| \( F_s \)   | Selected feature set        |
| \( CL \)    | Classification label        |

| notation    | annotation                  |
|-------------|----------------------------|
| \( D \)     | Data set                    |
| \( D' \)    | Data set after dimensionality reduction |
| \( F_0 \)   | Original feature set        |
| \( F_s \)   | Selected feature set        |
| \( CL \)    | Classification label        |
4.1 Feature Selection

Feature selection is one of the methods of dimensional reduction, and its goal is to find the optimal subset from the original feature sets. It eliminates irrelevant features, makes the dataset have fewer features, improves model accuracy and interpretability, and reduces runtime. It is an effective preprocessing method before high-dimensional data analysis. In order to reduce the computational cost, this paper uses an information theory method to first select a feature subset that express the nature of the data from the original features for clustering.

KL divergence, also known as information gain. Equation 2 can be understood as how much the information amount is reduced by using the distribution \( g(x) \) to approximate the unknown distribution \( f(x) \). Here, feature selection is performed by calculating the mutual information of each feature and classification label. Derived by formula (5), we get

\[
I(CL; f_i) = H(CL) - H(CL \mid f_i)
\]

\[
= - \sum_{c \in CL} p(c \mid f_i) \log p(c \mid f_i) \cdot \sum_{i} p(f_i) \sum_{c \in CL} p(c \mid f_i) \log p(c \mid f_i)
\]

\[
(6)
\]

Mutual information is a measure of the degree of dependence of two variables [8]. It is known from Eq. (5) that it determines the similarity of the joint distribution of the two variables and the product of the edge distribution. We use mutual information to measure the relationship between two variables. If they are closer, the value of mutual information is larger. Therefore, through mutual information calculation between features, we can select the required number of features to form a new feature subset \( F_S \).

The pseudo code of the related method is as follows:

**Algorithm 1: Feature selection based on information theoretic**

| Input: \( D \) | Output: \( D' \) |
|----------------|-----------------|
| 1: Initialize: \( F_S = \emptyset \); |  |
| 2: For each feature vector \( f_i \) do | \( I(c; f_i) = H(c) - H(c \mid f_i) \); |
| \( F_i^* = \arg \max_{f_i \in F_0} I(c; f_i) \); |  |
| 4: \( F_0 \leftarrow F_0 \setminus \{ f_i^* \} \); |  |
| 5: Repeat |  |

\[ C = \{ c_1, c_2, \ldots, c_k \} \]  Cluster
4.2 K-Means clustering

The main idea of K-Means is to first select K objects from the data set as the initial centroid, then calculate the distance from each object to the centroid and assign it to the cluster containing the nearest centroid; for each instance assigned, the centroid is recalculated based on the existing objects in the cluster. Repeat this process until all data objects are reassigned to the clusters, which is an iterative solution method. The algorithm is simple and easy to implement. It not only has a good clustering effect, but also the algorithm converges quickly. So K-means has many practical applications. However, the algorithm needs to preset the k, which is the number of clusters, and it is sensitive to the initial centroid. Different K and initial clusters centroid may lead to different results.

**Algorithm 2: K-Means clustering**

| Input: | \( D', K \) |
|---|---|
| Output: | \( C = \{C_1, C_2, ..., C_k\} \) |
| 1: Initialize: | \( C = \{\phi\} \), \( K = 1 \); |
| 2: Select the K points with the lowest similarity to other points from the objects as the initial centroid \( \mu_1, \mu_2, ..., \mu_k \); |
| 3: Grouping of data points by similarity measures |
| 4: Recalculate the centroid after each division |
| 5: Repeat this process until the centroid is unchanged |
| 6: Calculate the KL divergence between clusters by Equation 2 |

4.3 Consensus clustering

In the K-Means algorithm, different points are randomly selected as the initial centroid, and clusters of different sizes may be generated. It can be seen that it is sensitive to the choice of the initial centroid. Fern et al. found through experiments that the greater the difference between the base clustering, the better the integration result [5]. Here, K-means is run multiple times, each run randomly selects the data points as the centroid, which will result in different clustering results. We try to consensus these clusters through the framework of information theory to get the final results.

The goal of cluster integration is to find a new consensus partition \( P^* \) of \( D \), making objects in the same cluster in \( P^* \) more similar to objects in the same cluster in other partitions. We can calculate the mutual information values of the two partitions by Eq. (5) to measure their relevance:
\[ I(p_i; p_j) = \sum_{i=1}^{k} \sum_{j=1}^{k} p(c_{i}, c_{j}') \log \frac{p(c_{i}, c_{j}')}{p(c_{i})p(c_{j}')} \]  

Because there are \( K \) clusters in each partition, the overall MI is used to measure the correlation between its partitions.

\[ P^* = \arg\max_{P^*} \sum_{j=1}^{k} \sum_{i=1}^{k} p(C_{i}, C_{j}') \log \frac{p(C_{i}, C_{j}')}{p(C_{i})p(C_{j}')} \]  

5. Experiment

The experiments were conducted considering five datasets, all of which obtained from the UCI Machine Learning Repository. Their detailed information are shown in table 2, they are both numerical datasets.

**Table 2. Datasets used in the experiment**

| Dataset                  | Objects | Attributes | Clusters |
|--------------------------|---------|------------|----------|
| Breast Cancer            | 286     | 9          | 2        |
| Wine Quality             | 178     | 13         | 3        |
| Soybean                  | 47      | 35         | 4        |
| Alcohol QCM Sensor       | 125     | 8          | 5        |
| Glass Identification     | 214     | 9          | 6        |

In order to verify the proposed method, we directly compared the method presented in this article with the K-means integration in the experiment of design. To ensure that the algorithm compares under the same conditions, the experiment did not consider the relationship between the K and the clustering result, but instead set \( K \) directly equals to the true number of clusters in each dataset. At the same time, for easy comparison, we have run the same number of times for both methods: the K-Means is run 50 times on \( D \), and then integrate the results; the same as the algorithm presented in this paper.

We use the Normalized Mutual Information (NMI) score to evaluate the clustering ensemble results. NMI is a commonly used clustering evaluation indicator, which can measure the correlation between vectors of label of the two clustering results, and ignore the permutations in labels. For the final clustering result \( P^* \) and the ground-truth clustering result \( P^{\text{gt}} \), the NMI is calculated [10] as:

\[ \text{NMI}(P^*, P^{\text{gt}}) = \frac{2I(P^*, P^{\text{gt}})}{H(P^*) + H(P^{\text{gt}})} \]  

where \( H(P^*) = \sum_{c} |c| \log \frac{|c|}{n} \), and \( H(P^{\text{gt}}) = \sum_{c} |c| \log \frac{|c|}{n} \).

\( \text{NMI}(P^*, P^{\text{gt}}) \in [0,1] \), the larger the NMI value, the more consistent \( P^* \) is with the ground-truth classification result \( P^{\text{gt}} \). The two algorithms then run 10 times each, and the average results are shown in Table 3.

**Table 3. The average performance measured by NMI**
6. Conclusion

Clustering research on high-dimensional data has important significance. In this article, we have proposed an information-theoretic based method for clustering high-dimensional data. First, we calculate mutual information indicators between features, select the essential features from the original data set, and then calculate the similarity of the data instances to select the K lowest similarity points as the initial centroids, and use K-Means method for clustering. The algorithm is run ten times, and each generating the same number of clusters; finally, we define the utility function to find the optimal partition, which is most similar to the ground-truth clustering result. Through comparison, we verify the effectiveness of the method, especially for larger dimensional data sets. In addition, clustering ensemble effectively overcomes the uncertainty of base clustering, and integrate the results of base clustering to improve the quality of clustering.

Subspace clustering technology selects different subspaces from high-dimensional spaces, and then performs clustering on the low-dimensional space. It is one of the very important research fields in high-dimensional data analysis. In future, our research will combine the advantages of subspace and clustering ensemble, and it is a valuable research direction to aggregate the results of subspace clustering.

References

[1] A P Topchy, A K Jain, W F Punch, Combining multiple weak clusterings. International conference on data mining, San Francisco 2003, 331-338.
[2] Caiming Zhong, Xiaodong Yue, Zehua Zhang, Jingsheng Lei, A clustering ensemble. Pattern Recognition, 2015, 48(8), 2699-2709.
[3] D Md Farid, A Nowe, B Manderick, A feature grouping method for ensemble clustering of high-dimensional genomic big data. Future technologies conference, 2016, 260-268.
[4] E Gokcay, J C Principe, Information theoretic clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence.2002, 2(24), 158-171.
[5] Fern X Z, Brodley C E. Random projection for high dimensional data clustering: A cluster ensemble approach, ICML.2003, 3:187-192.
[6] I S. Dhillon, S Mallela, R Kumar, A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification. Journal of Machine Learning Research, 2003.3: 1265-1287.
[7] Juhua Hu, Jian Pei, Subspace Multi-Clustering: A Review. Knowledge and Information Systems. 2018, 56(2), 257-284.
[8] M Bennasar, Y Hicks, R Setchi, Feature selection using Joint Mutual Information Maximisation. Expert Systems with Applications, 2015, 42(22), 8520-8532.
[9] T Alqurashi, Wenjia Wang, Clustering ensemble method. International Journal of Machine Learning and Cybernetics, 2019(10):1227-1246.
[10] Wei Liang, Yuanjian Zhang, Jianfeng Xu, Deyu Lin, Optimization of Basic Clustering
for Ensemble Clustering: An Information-Theoretic Perspective. IEEE Access. 2019(7), 179048-179062.

[11] Xuan Huang, Lei Wu, Yinsong Ye. A Review on Dimensionality Reduction Techniques. International Journal of Pattern Recognition and Artificial Intelligence. 2019, 33(10). 1950017.