

Review of Modern Techniques of Qualitative Data Clustering

Sergey Cherevko and Andrey Malikov

The North Caucasus Federal University, Institute of Information Technology and Telecommunications

cherevkosa92@gmail.com, malikov@ncstu.ru

Abstract: This article made a brief comparative survey of modern clustering algorithms quantitative and qualitative data. As a practical component to the process of analysis of algorithms used in the task of analyzing the consumer basket. The examples of commercial use of clustering algorithms, described the current problems of using cluster analysis.

Keywords: cluster analysis, qualitative data, algorithm Clope, Kohonen's maps, overview of clustering algorithms

1 Introduction

In general, the concept of clustering and cluster analysis can be described as segmentation of a set of objects into different groups, called clusters. Affinity between objects from a same cluster should be higher than affinity between objects from different clusters. In the process of applying the techniques of clustering in the bulk of the tasks the number of clusters is unknown in advance - this characteristic is defined in the algorithm, but it's worth noting that there are algorithms that require initial quantification of finite groups.

The possibility of applying different clustering algorithms and their performance is determined by the following set of indicators:

1. Type of data to be clustered. Conditionally data types are divided into numerical continuous (values of a certain period: adult height, projectile range), discrete numerical (values from a list of some specific numbers: the number of children in the family, the number of clients seeking day) and categorical (descriptive characteristic feature of any object).
2. Dimensionality of the data set to be clustered. In accordance with this feature, you can select the algorithms that work only with sets of small sizes (for large samples, the effectiveness of these techniques falls due to low performance of the algorithm), and algorithms designed to handle large and very large data sets (in this case, however, quality of created clusters may suffer) [1].
3. Dimensionality of the data set being evaluated. If the set contains a large amount of emissions, some algorithms can give incorrect results misrepresent the nature and composition of the clusters.

2 Descriptive Comparison of Clustering Algorithms

If we talk about the current level of clustering techniques, it should be noticed that the range of techniques regarded to the analysis of quantitative data is wide enough and allows you to perform the tasks of clustering in accordance with all requirements to processing speed and quality of the final result. The most well-known representatives of numeric data clustering algorithms are:

1. Algorithm Clustering Using Representatives.
2. Algorithm Balanced Iterative Reducing and Clustering using Hierarchies.
3. Algorithm HCM (Hard C - Means).

Usage of the aforementioned methods for clustering categorical data is inefficient, and often impossible. The main difficulties are associated with high dimensionality and huge volumes of data that often characterizes such databases, because the pairwise comparison of the characteristics of objects from multi-million records database tables can take quite a long period of time. Clustering algorithms of aforementioned types are currently having a number of special requirements - recommendations that will optimize performance when working with large amounts of data:

1. Ability to work in a limited amount of RAM.
2. The algorithm should work under the condition that the information from the database can be gathered only in the forward-only cursor.
3. Ability to abort the algorithm with preservation of intermediate results with the possibility of resuming the process of data processing.
4. Minimizing the number of requests for the full database table scan [2].

Currently, the most promising and popular clustering algorithms qualitative data include:

Currently, the most promising and popular algorithms for clustering qualitative data include:

1. Algorithm Clope. The main advantages of this algorithm are speed and quality of clustering which are achieved by using the estimated global optimization criterion based on the maximization of the gradient histogram cluster height. During its operation, the algorithm stores a small amount of information on each of the clusters in memory, and requires a minimum number of data sets scans. Number of clusters is automatically selected by the algorithm based on the coefficient of repulsion, which was originally set by the user (the more the level of the coefficient is, the lower the level of similarity will be, and more clusters will be generated). [3]

2. Algorithm LargeItem, which was developed in 1999 as an optimized algorithm for clustering data sets based on the appraised, absolute function that uses the support parameter.
3. Kohonen maps (this method is applicable to both qualitative and quantitative indicators). Allows you to identify useful and non-trivial patterns, consider the influence of hundreds of factors, visualize complex multidimensional clusters in the form of clear and accessible maps.
4. Hierarchical clustering, as an example of MST (Algorithm based on Minimum Spanning Trees). MST algorithm first builds minimum spanning tree on the graph, and then sequentially removes edge with the largest weight. The disadvantage is the high degree of dependence on emissions contained in the data sets [4].

3 Comparison of Algorithms Clope and Kohonen's Maps for the Problem of the Consumer Basket

For a more practical comparison of algorithms we will describe a typical clustering problem: the consumer's basket analysis. The aim is to define a set of products that are most often found in one check (or order).

3.1 Algorithm Clope

"Clope" algorithm functionality is based on the idea of maximizing the global cost function, which increases the degree of similarity of transactions in the clusters by increasing the parameter of cluster histogram. Consider a simple example, we have 5 consumer checks:

1. {[bread, milk]}.
2. {[bread, milk, sour cream]}.
3. {[bread, sour cream, cheese]}.
4. {[cheese, sausage]}.
5. {[cheese, sausage, peppers]}.

Suppose that we already have two partitions into clusters:

1. {[bread, milk, bread, milk, sour cream, bread, sour cream, cheese] [cheese, sausage, cheese, sausage, peppers]}.
2. {[bread, milk, bread, milk, sour cream]}, {[bread, sour cream, cheese, cheese, sausage, cheese, sausage, peppers]}.

Let's calculate the height (H) and width of the cluster (W) for both clusters.

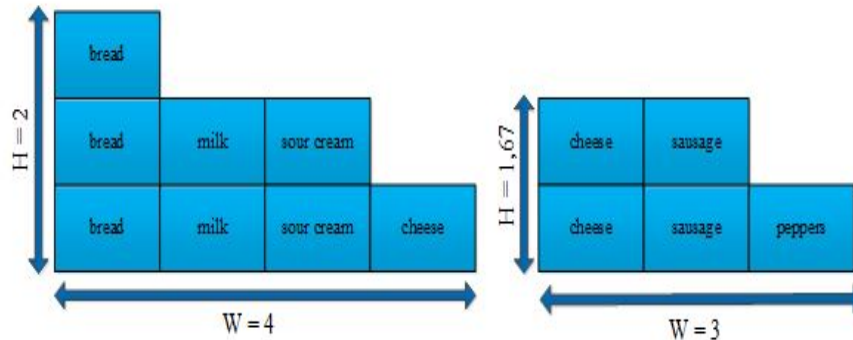


Figure 1 - distribution of the first cluster

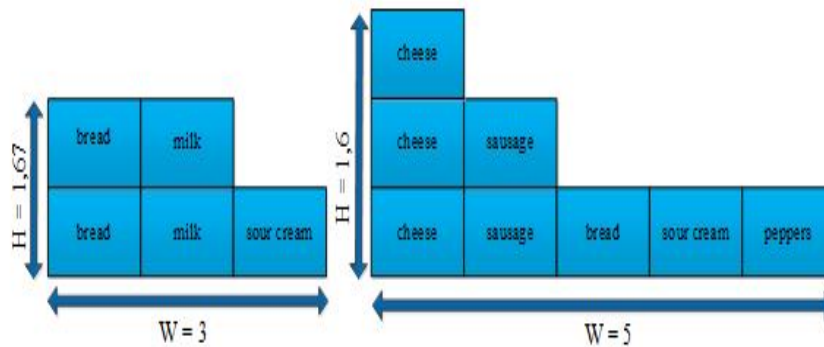


Figure 2 - distribution of the second cluster

H - height of the chart, is calculated as the average height (density) of all the columns in the chart. For example, for the first cluster we got $H = 2$ as a result of the following arithmetic operation:

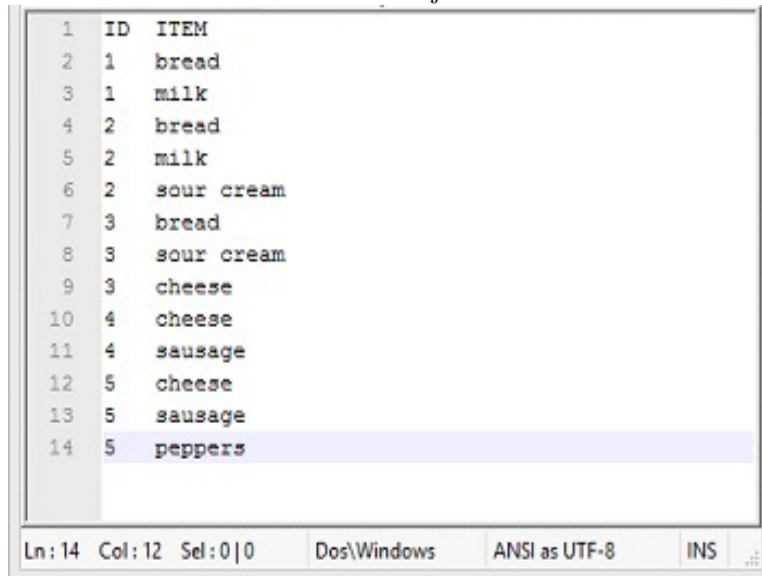
$$(3 + 2 + 2 + 1) / 4 = 2.$$

W - the width of the cluster, which is equal to the number of columns in the cluster. To determine the quality of the partition clusters by Clope it is necessary to calculate cluster diagram parameter (H / W) for both distributions. For the first cluster it will be $(2 + 1.67) / (3 + 4) = 0.52$. For the second cluster, respectively $(1.67 + 1.67) / (3 + 5) = 0.41$. Obviously, the first partition is better option because cluster diagram parameter is higher, which indicates that the transaction have a large overlap with each other.

3.2 Kohonen's Maps

Let's try to perform this distribution using Kohonen maps. Compose a text file containing information about the checks. Column "Id" corresponds to the

check box and "Item" describes the subject contained in the check.



1	ID	ITEM	
2	1	bread	
3	1	milk	
4	2	bread	
5	2	milk	
6	2	sour cream	
7	3	bread	
8	3	sour cream	
9	3	cheese	
10	4	cheese	
11	4	sausage	
12	5	cheese	
13	5	sausage	
14	5	peppers	

Figure 3 - The composition of a text file for processing

Processing options in the input file define that we want to get 2 clusters as a result of processing, composed on the basis of field "Item". Neurons odds left by default, without changing the coefficients. The result of the processing is following (see Figure 4, 5).

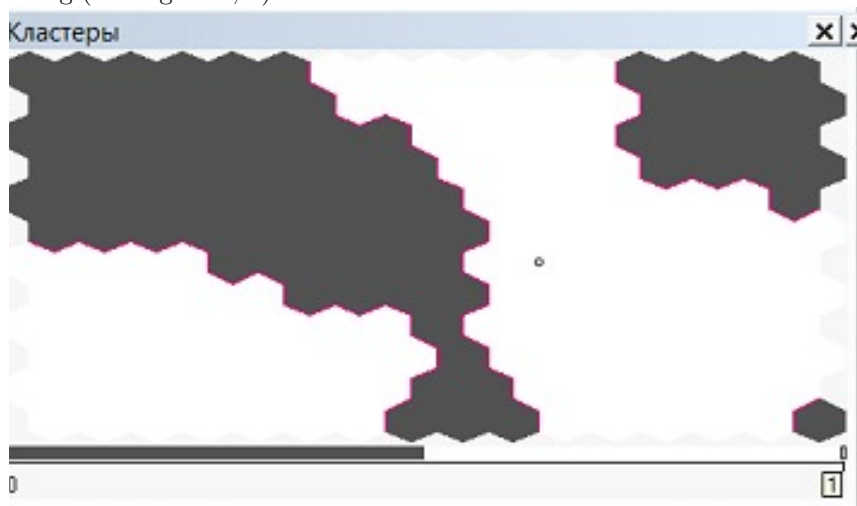


Figure 4 - a graphic representation of the clusters

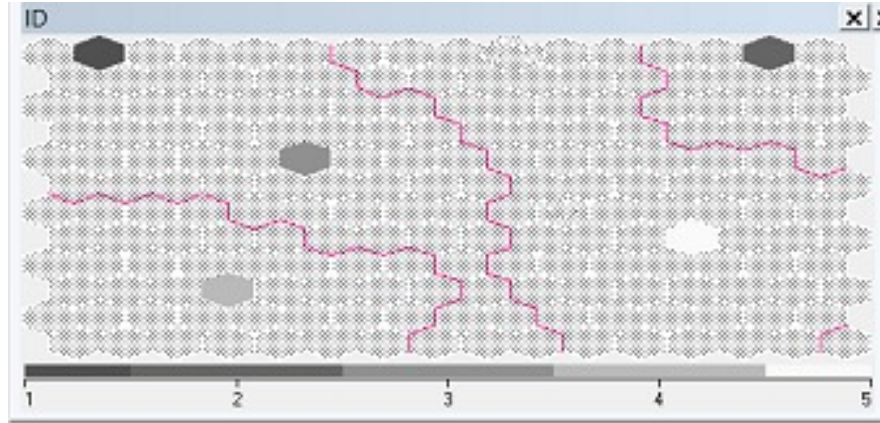


Figure 5 - Distribution of the consumer basket items in clusters

According to the results of distribution two clusters were allocated, the first checks were 1,2,3, the second 4,5. Distribution is obtained similarly to distribution provided by Clope algorithm. Basic formula, in which the distribution is a function of the neighborhood, allows us to define a "measure neighborhood" nodes compared:

$$h_{ci}(t) = \alpha(t) \times \exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right)$$
, where $0 < \alpha(t) < 1$ is learner and decreasing factor, r_i and r_j are coordinates of nodes $M_i(t)$ and $M_c(t)$. The last step of the algorithm is to change the vectors of weights on approaching the observation under consideration:

$$m_i(t) = m_i(t-1) + h_{ci}(t) \times (x(t) - m_i(t-1)).$$

As the results of the algorithm, the user has a visual map, where similar indicators will be grouped into visually isolated clusters.

4 Conclusion

Nowadays there is enough of active practical usage of clustering methods, as it makes work with homogeneous data sets much easier and allows the usage of specific methods of treatment which are not suitable for the total sample. In fact, the bigger and bigger funds are conducted by different companies around the globe to provide scientific segmentation and clustering of customers and end-products. The most famous events in this direction are: "Svyaznoi" company with their program "Svyaznoi Club", Bank of Moscow with the targeted marketing automation system SAS Marketing Automation, "PSB", "Tinkoff" bank and many others. For example, the company "Svyaznoi" in its loyalty program "Svyaznoi Club" implemented deep customer segmentation - cardholders of "Svyaznoi Club". Specialists from BaseGroup Labs and analytical platform Deductor were involved to solve this problem. They helped to process the entire array of raw data, and built more than 120 classification models for client's response per share. Despite the fairly extensive use of qualitative data clustering

methods there are a number of topical issues that hinder the work and further developments: 1. Lack of comparable treatment for describing objects with attributes of different nature, measured in different units. 2. The complexity of making a formalized description of objects having both qualitative and quantitative traits. 3. The problem of detecting and displaying the artifacts in the source data. The need to develop such clustering method that would identify artifacts in the source data and either exclude them from the final result, or carry specific processing to visually represent the fact of "break out".

References

- [1] Article Ke Wang, ChuXu, Bing Liu Clustering Transactions Using Large Items 2003
- [2] Yang, Y., Guan, H., You. J. CLOPE: A fast and Effective Clustering Algorithm for Transactional Data In Proc. of SIGKDD'02, July 23-26, 2002, Edmonton, Alberta, Canada.
- [3] Classification and comparison of clustering methods / IM transmission line protected [electronic resource].
- [4] F. Hooper, F.Klawonn, R.Kruse, T.Runkler, Fuzzy Clustering. Chichester, United Kingdom: Wiley, 1999