A NEW APPROACH TO REDUCTION OF LARGE DATA SET

D.A. Viattchenin
United Institute of Informatics Problems of the NAS of Belarus
Minsk, Belarus
e-mail: viattchenin@mail.ru

Abstract

The paper deals with the problem of the large data reduction. The proposed approach is based on reduction of the number of objects in the initial large data set and representation of each new object from the reduced data set as a vector of intervals. An illustrative example is given and preliminary conclusions are formulated.

1 Introduction

In applied statistics the large data usually represent a high-dimensional data sets as well as the data sets with a large number of objects. The large data usually include data sets with sizes beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time.

Given the ubiquity of large high-dimensional data sets, and the need not only to transmit, archive, and reduce them, but also to represent and analyze their content, corresponding tools for the computational dissection, analysis and representing of complex datasets whose size defies simplistic analysis must be developed. These tools should include a representation which allows exploitation of nonlinearity, supports fast multiresolution algorithms, incorporates a priori information and constraints, and provides for flexibility of adaptation to the application.

The large data sets are difficult to work with for several reasons. They are difficult to visualize, and it is difficult to understand what sort of errors and biases are present in them. They are computationally expensive to process, and often the cost of learning is hard to predict – for instance, and algorithm that runs quickly in a dataset that fits in memory may be exorbitantly expensive when the dataset is too large for memory.

The most common methods of the data reduction are methods to reduce the dimensionality of attributes, such as principal component analysis and multidimensional scaling. However, the problem of reduce the number of objects in the studied data set is also very actual, and exclusion of repeated objects is the most common approach to solving the problem. A new approach to the problem solving is presented in the paper and the approach is based on representation of the initial large data set by the data set with a smaller number of observations, where each object is described by the vector of intervals. In other words, a value of some attribute for an object should be considered as an interval of values. So, a problem of the interval-valued data processing is arises.

2 Heuristic for Reduction of Large Data

The object data clustering methods can be applied if the objects are represented as points in some multidimensional space $I^{m_1}(X)$. In other words, the data which is composed of n objects and m_1 attributes is denoted as $\hat{X}_{n\times m_1} = \left[\hat{x}_i^{t_1}\right]$, $i = 1, \ldots, n$, $t_1 = 1, \ldots, m_1$ and the data are called sometimes the two-way data [4]. Let $X =$ $= \{x_1, \ldots, x_n\}$ is the set of objects. So, the two-way data matrix can be represented as follows:

$$\hat{X}_{n\times m_1} = \begin{pmatrix} \hat{x}_1^1 & \hat{x}_1^2 & \ldots & \hat{x}_1^{m_1} \\ \hat{x}_2^1 & \hat{x}_2^2 & \ldots & \hat{x}_2^{m_1} \\ \ldots & \ldots & \ldots & \ldots \\ \hat{x}_n^1 & \hat{x}_n^2 & \ldots & \hat{x}_n^{m_1} \end{pmatrix}. \tag{1}$$

So, the two-way data matrix can be represented as $\hat{X} = \{\hat{x}^1, \ldots, \hat{x}^{m_1}\}$ using n-dimensional column vectors $\hat{x}^{t_1}$, $t_1 = 1, \ldots, m_1$, composed of the elements of the t_1-th column of $\hat{X}$.

Let us assume that the value $n = card(X)$ is very large. The matter of the proposed method can be formulated as follows: the initial large data set must be pre-processed by some clustering procedure for some value $\tilde{n} << n$ and the reduced data set $\tilde{X} =$ $= \{x_1, \ldots, x_{\tilde{n}}\}$ must be constructed. The reduced data set $\tilde{X} = \{x_1, \ldots, x_{\tilde{n}}\}$ described by m_1 interval attributes $\left\{\tilde{\tilde{x}}^1, \ldots, \tilde{\tilde{x}}^{m_1}\right\}$. An interval attribute $\tilde{\tilde{x}}^{t_1}$ is a correspondence defined from X in $\mathfrak{R}$ such that for each $x_{\tilde{i}} \in X$, $\tilde{\tilde{x}}_{\tilde{i}}^{t_1} = \left[\tilde{\tilde{x}}^{t_1(min)}, \tilde{\tilde{x}}^{t_1(max)}\right] \in \mathfrak{J}$, where $\mathfrak{J} = \left\{\left[\tilde{\tilde{x}}^{t_1(min)}, \tilde{\tilde{x}}^{t_1(max)}\right] : \tilde{\tilde{x}}^{t_1(min)}, \tilde{\tilde{x}}^{t_1(max)} \in \mathfrak{R}, \tilde{\tilde{x}}^{t_1(min)} \leq \tilde{\tilde{x}}^{t_1(max)}\right\}$ is the set of closed intervals defined from $\mathfrak{R}$ [2]. In other words, each object $x_{\tilde{i}} \in \tilde{X}$ is represented as a vector of intervals $x_{\tilde{i}} = \left(\tilde{\tilde{x}}_1^{t_1}, \ldots, \tilde{\tilde{x}}_{\tilde{n}}^{t_1}\right)$, where $\tilde{\tilde{x}}_{\tilde{i}}^{t_1} = \left[\tilde{\tilde{x}}^{t_1(min)}, \tilde{\tilde{x}}^{t_1(max)}\right] \in \mathfrak{J}$. So, the reduced data $\tilde{X} = \{x_1, \ldots, x_{\tilde{n}}\}$ are an interval-valued data and the data table $\tilde{\tilde{X}}_{\tilde{n}\times m_1} =$ $= \left[\tilde{\tilde{x}}_{\tilde{i}}^{t_1}\right]$ is made up of $\tilde{n}$ rows representing the $\tilde{n}$ objects and m_1 columns representing m_1 interval attributes.

The value $\tilde{n} << n$ must be discovered. For the purpose, a heuristic formula

$$\tilde{n} = \lfloor\sqrt{n}\rfloor = \lfloor\sqrt{cardX}\rfloor, \tag{2}$$

can be used. In general, the proposed method for the data reduction can be summarized as a procedure as given below:

1. The initial data set $X = \{x_1, \ldots, x_n\}$ must be processed some clustering procedure for $\tilde{n}$ classes and corresponding hard partition $\tilde{X}$ will be obtained;

2. Calculate values $\tilde{\tilde{x}}^{t_1(min)}$ and $\tilde{\tilde{x}}^{t_1(max)}$ of attributes $\hat{x}^{t_1}$, $t_1 = 1, \ldots, m_1$ for each class $x_{\tilde{i}} \in \tilde{X}$, $\tilde{i} = 1, \ldots, \tilde{n}$;

3. Construct the set of vector of intervals $x_{\tilde{i}} = \left(\tilde{\tilde{x}}_1^{t_1}, \ldots, \tilde{\tilde{x}}_{\tilde{n}}^{t_1}\right)$, $\tilde{i} = 1, \ldots, \tilde{n}$, where $\tilde{\tilde{x}}_{\tilde{i}}^{t_1} = \left[\tilde{\tilde{x}}^{t_1(min)}, \tilde{\tilde{x}}^{t_1(max)}\right]$.

So, the reduced data $\tilde{X} = \{x_1, \ldots, x_{\tilde{n}}\}$ are the interval-valued data and the data set can be processed using some clustering technique [3], [5].

3 An Illustrative Example

The proposed approach to the data reduction must be explained by simple example and the Anderson's iris data [1] can be used for the purpose. The Iris database is the most known database to be found in the pattern recognition literature. The data set represents different categories of Iris plants having four attribute values. The four attribute values represent the sepal length, sepal width, petal length and petal width measured for 150 irises. It has three classes Setosa, Versicolor and Virginica, with 50 samples per class. According to formula (2) we obtain $\tilde{n} = \lfloor \sqrt{150} \rfloor = 12$. So, the iris data set was classified using k-means method where the number of classes was equal 12. The reduced data set is presented by Table 1.

Table 1. The reduced Anderson's iris data set

Numbers	Attributes			
of objects	Sepal length	Sepal width	Petal length	Petal width
1	[7.1, 7.9]	[2.6, 3.8]	[5.8, 6.9]	[1.6, 2.5]
2	[5.6, 6.3]	[2.2, 3.2]	[4.8, 5.1]	[1.5, 2.4]
3	[6.1, 6.7]	[2.5, 3.1]	[5.2, 5.8]	[1.4, 2.2]
4	[4.8, 5.5]	[3.3, 3.8]	[1.3, 1.9]	[0.1, 0.6]
5	[6.1, 7.0]	[2.8, 3.4]	[4.3, 5.0]	[1.2, 1.7]
6	[4.3, 5.0]	[2.3, 3.6]	[1.0, 1.6]	[0.1, 0.3]
7	[6.2, 6.9]	[3.0, 3.4]	[5.1, 6.0]	[2.0, 2.5]
8	[6.0, 6.3]	[2.2, 2.8]	[4.0, 4.5]	[1.0, 1.5]
9	[5.2, 5.8]	[2.3, 2.9]	[3.5, 4.1]	[1.0, 1.4]
10	[4.9, 6.0]	[2.5, 3.0]	[4.1, 4.5]	[1.2, 1.7]
11	[5.2, 5.8]	[3.7, 4.4]	[1.2, 1.7]	[0.1, 0.4]
12	[4.9, 5.1]	[2.0, 2.5]	[3.0, 3.5]	[1.0, 1.1]

So, each cell of this table contains an interval $\tilde{\hat{x}}_{\tilde{i}}^{t_1} = \left[\tilde{\hat{x}}^{t_1(min)}, \tilde{\hat{x}}^{t_1(max)}\right]$, $\tilde{i} \in \{1, \ldots, 12\}$, $t_1 = 1, \ldots, 4$ and the corresponding interval data matrix $\tilde{\hat{X}}_{12\times4} = \left[\tilde{\hat{x}}_{\tilde{i}}^{t_1}\right]$ can be processed by some clustering technique.

4 Conclusions

The method for reduction of large data sets is proposed in the paper. The method based on preliminary clustering of the initial large data set using some fast procedure for following representation the reduced data set by some interval data matrix for following processing by precise clustering procedure. An experiment with the Anderson's iris data set shows the usefulness and effectiveness of the proposed method.

References

[1] Anderson E. (1935). The Irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*. Vol. **59**, pp. 2-5.

[2] Bock H.H., Diday E. (2000). *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Heidelberg.

[3] de Carvalho F.d.A.T. (2007). Fuzzy C-Means Clustering Methods for Symbolic Interval Data. *Pattern Recognition Letters*. Vol. **28**, pp. 423-437.

[4] Sato-Ilic M., Jain L.C. (2006). *Innovations in Fuzzy Clustering: Theory and Applications*. Springer-Verlag, Heidelberg.

[5] Viattchenin D.A. (2013). *A Heuristic Approach to Possibilistic Clustering: Algorithms and Applications*. Springer-Verlag, Heidelberg.