System for a cluster analysis

Yuri Stekh, Fajsal M.E. Sardieh, Mykhaylo Lobur

Abstract - **This paper focuses on structure of dialog graphical system for pattern recognition with help of distance function. In this paper is evaluate the performance of different criterion functions and algoritms for the problem of clustering large datasets.**

Keywords – **Clustering algorithm, Criterion function, large datasets, distance function.**

The set of programs of the graphic interactive system of patterns recognition based on the function of distance is developed. Pattern is presented by a point in n - measured space.

$$\bar{x} = (x_1, x_2, ..., x_n) \quad (1)$$

where $\bar{x}$ – vector of pattern's properties,

$x_1, x_2, ..., x_n$ – pattern's properties.

The large dataset of documents may be used as object of investigation. each document is represented by the term-frequency (TF) vector

$$d_{tf} = (tf_1, tf_2, ..., tf_m) \quad (2)$$

where tf_i is the frequency of the *i*th term in the document. A widely used refinement to this model is to weight each term based on its inverse document frequency (IDF) in the document collection. The motivation behind this weighting is that terms appearing frequently in many documents have limited discrimination power, and for this reason they need to be de-emphasized. This is commonly done by multiplying the frequency of each term *i* by $\log\left(N/df_i\right)$, where N is the total number of documents in the collection, and df_i is the number of documents that contain the ith term (*i.e.*, document frequency). This leads to the *tf-idf* representation of the document, *i.e.*

$$d_{tfidf} = \begin{pmatrix} tf_1 \log\left(N/df_1\right), tf_2 \log\left(N/df_2\right), ... \\ , tf_m \log\left(N/df_m\right) \end{pmatrix} \quad (3)$$

To account for documents of different lengths, the length of each document vector is normalized so that it is of unit length $\left(\left\|d_{tfidf}\right\| = 1\right)$ that is each document is a vector in the unit hypersphere. In the rest of the paper, we will assume that the vector representation for each document has been weighted using *tf-idf* and it has been normalized so that it is of unit length.

For the calculation of distances between patterns Euclid distance is used:

$$dist(\overline{x,y}) = \sqrt{\sum (x_i - y_i)^2} \quad (4)$$

square of Euclid distance

$$dist(\overline{x,y}) = \sum (x_i - y_i)^2 \quad (5)$$

Chebyshev distance

$$dist(\overline{x,y}) = \sum |x_i - y_i| \quad (6)$$

power distance

$$dist(\overline{x,y}) = \left(\sum |x_i - y_i|^p\right)^{1/r} \quad (7)$$

angle distance

$$dist(\overline{x,y}) = \frac{x' y}{\|x\| \|y\|} \quad (8)$$

Mohalonabis distance.

$$dist(\overline{x,y}) = (x-y)' C^{-1} (x-y) \quad (9)$$

The set of programs includes the followings algorithms: the algorithm of threshold size, maxmin algorithm, k-means algorithm, algorithm of PAM (partitioning around medios), ISODATA algorithm (Iterative Seif-organizing Data Analysis Techniques) and BIRCH algorithm (Balanced Iterative Clastering using of Hierarchies).

The system contains a comfortable graphical user interface, provides input, editing of initial data, storing data in files and loading from files, choice of algorithm of calculation, implementation of algorithm in the automatic or step mode, displaying the results in the character or graphic modes. Input data can be in the form of array of 2D or 3D representation of points. Displaying of results is fulfilled in 2D or 3D modes.

REFERENCES

[1] Дж. Ту, Р. Гонсалес "Принципы распознавания образов" М., Мир, 1978.

[2] Батыршин И.З., Хабибулин Р.Ф. Тестирование кластерных алгоритмов на инвариантность относительно нумерации объектов. - Известия академии наук. Теория и системы управления.- 1997, 2.

[3] Батыршин И.З., Хабибулин Р.Ф. Разработка алгоритмов когнитивного кластерного анализа, в кн.: Обработка текста и когнитивные технологии, вып. 3/Под ред. Соловьева В.Д. - Пущино, 1999.

[4] Ермаков А.Е. Тематический анализ текста с выявлением сверхфразовой структуры // Информационные технологии. - 2000. - N 11.

[5] Yu. Stekh "*Two clastering algorithms for large datasets*" in *Proceeding of the XIV Ukrainian* – Polish Conference CADMD'2006, Poljana, Ukraine, May 22-23, 2006.

Yuri Stekh, Fajsal M.E. Sardieh, Mykhaylo Lobur – CAD/CAM Department, Lviv Polytechnic National University, 12, S. Bandera Str., Lviv, 79013, UKRAINE,
E-mail: yuristekh@yahoo.com