UDC 620.179.16

S.A. Babichev

APPLICATION OF ENTROPY CRITERION FOR AN ESTIMATING OF QUALITY OF DNA MICROARRAY DATA NORMALIZATION

In the article the investigation on use of Shannon entropy criterion for an estimation of quality of DNA microarray data normalization is presented. The methods of linear and nonlinear normalization and contrast method were used as normalization methods. It is shown the best quality of normalization achieved by using the contrast method, that proved by a minimum value of the Shannon entropy.

Keywords: DNA microarray, data normalization, the Shannon entropy criterion.

1. Introduction.

Normalization is one of the major stages of data preprocessing in which empirical data are transformed to the same range that allow to carry out their comparative analysis. Furthermore, the character of distribution of data in the chosen range of values has been changing that defines quantity and quality of evaluating information. The character distribution of data received in the process of normalization is defined by a kind of used transfer function. Quality of data preprocessing in many respects depends on a choise of function of normalization what in turn defines quality of the preprocessing information. Questions of normalization DNA microarray data in enough detal were discussed in [1]. According to character of distribution data existing methods of normalization can be divided into two subgroups: the methods using a basic array of genes and methods using all set of investigated data. The first groups of methods are in turn subdivided into linear and nonlinear. The second groups of methods include methods of cyclic loes, contrast based method and quantile normalization [2,3]. Each of methods has advantages and disadvantages. In [2] it is shown that with other things being equal better works a method quantile normalization entering as base method in a complex data preprocessing RMA. However, it is necessary to notice that the evaluation of quality of data normalization was mainly made by visual inspection according to degree of variability of data on appropriate plots of distribution without use of any quantitative criteria. This approach is subjective, because the normalized data can contain the latten laws invisible on schedules of corresponding dependences. In

this paper it is offered to use along with the traditional methods of visual an estimation of quality normalization quantitative criterion of entropy [4,5], that will allow to raise objectivity of an estimate and to choose for current data more optimal method of normalization.

In this case as entropy we will understand a quantitative measure of ordering of structural elements in system. According to statistical definition of entropy:

$$S = k \cdot \ln W \tag{1}$$

where k - Boltzmann constant, W - probability of thermodynamic conditions defined as quantity of possible microconditions allowing to realize given macroconditions.

According to formula (1) the system in which all structural elements are ordered, has minimum entropy, but thus we have maximum information about structure of the system and configurations of its components. At increase of the level of disorder of the elements in the system (presence of noise component), entropy of the system increases because quantity of microcondition realizing given macrocondition increases, however objective information about true condition of the structural elements in the system decreases. K. Shannon proposed the formula [6] allowing to estimate information's entropy of an investigated signal:

$$H = -K \cdot \sum_{i=1}^{n} p_i \log(p_i) \tag{2}$$

where K – the positive constant entered in the equation for coordination of dimension, p_i - the probability of the i-th event. At DNA microarray analysis the event is understood as intensity of light in the given point. If to assept that s_i - the level of intensity in i-th point, the probability of realization of the given condition is connected with the function of realization of the given condition by equation [7]:

$$p_i = s_i^2 \tag{3}$$

Thus, (2) taking into account (3) it is possible to present formulay as follows:

$$H = -K \cdot \sum_{i=1}^{n} s_i \log(s_i) \tag{4}$$

 ISSN 1562-9945

According to the received formula it is possible to draw a conclusion that in process of increases of quality of data normalization the quantity of information about a useful component of the signal increases too that agreement increases signal-noise-relative. The value of Shannon entropy criterion at the given case will seek to minimum.

The purpose of this article is to develop the system of an estimation of quality of DNA microarray data normalization on the basis of quantitative criterion of entropy.

2. Results

Suppose that array of experimental data obtained as a result of scanning of DNA microarray is presened by the matrix dimension $n \times m$:

$$I = \begin{cases} x_{11}, x_{12}, \ldots, x_{1m}; \\ x_{21}, x_{22}, \ldots, x_{2m}; \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_{n1}, x_{n2}, \ldots, x_{nm}. \end{cases} \quad (5)$$

where n – a number of observed objects or a number of genes which expression level is necessary to be determined; m – a number of features that are characteristics of any object. Each element of the matrix (5) represents intensity of light corresponding of the given gene. For comparative analysis objects of the matrix for purpose of classification it is necessary to normalize, i.e. their value transform to the same range. Normalization of data was performed using the methods of linear and nonlinear normalization with application of the logistic transfer function and the contrast method.

Linear normalization algorithm supposed the following steps:

–from the set of vectors of matrix (5) it is selected the basis vector for which the mean value is calculated:

$$\overline{x}_{bs} = \frac{\sum_{i=1}^{m} x_{bsi}}{m} \quad (6)$$

–for each vector of array the trimmed mean is calculated herewith discarded 2% minimum and maximum values:

$$\overline{x}_k = \frac{\sum_{i=1}^{0,98m} x_{ki}}{0,98m} \quad (7)$$

–normalizing values for each vector of array are calculated accordingly equation:

$$x'_{ki} = \frac{\overline{x}_{bs}}{\overline{x}_k} x_{ki} \quad (8)$$

Nonlinear normalization was performed using logistic transfer function as follows:

–mean value of corresponding vector of array is calculated:

$$\overline{x}_k = \frac{\sum_{i=1}^{m} x_{ki}}{m} \quad (9)$$

–standard deviation is calculated for each vector of array:

$$\sigma_k = \sqrt{\frac{\sum_{i=1}^{m} (x_{ki} - \overline{x}_k)}{m \cdot (m-1)}} \quad (10)$$

–vector of arguments of the logistic function is calculated using formula:

$$a_{ki} = \frac{x_{ki} - \overline{x}_k}{\sigma_k} \quad (11)$$

–normalized values for each vector of array are calculated:

$$x'_{ki} = \frac{1}{1 + e^{-a_{ki}}} \quad (12)$$

The contrast method supposes that between vectors M and A prevail linear regression dependency. Vectors M and A represent as a set of the following values:

$$M_{ki} = \log_2\left(\frac{x_{ki}}{x_{bsi}}\right) \quad (13)$$

$$A_{ki} = \log_2(x_{ki} \cdot x_{bsi}) \quad (14)$$

 ISSN 1562-9945

where x_k - k-th vector of investigation dataset; x_{bs} - basis vector of the given array.

Algorithm of the contrast method supposes the following steps:

–select a basis vector;

–determine parameters M and A for each investigated vector of data array accordingly equations (13) and (14);

–calculate normalizing correction for each element of the corresponding vector:

$$\delta M_{ki} = M_{ki} - \hat{M}_{ki} \tag{15}$$

where $\hat{M}_{ki}$ - value of regression function corresponding to i-th sample of k-vector;

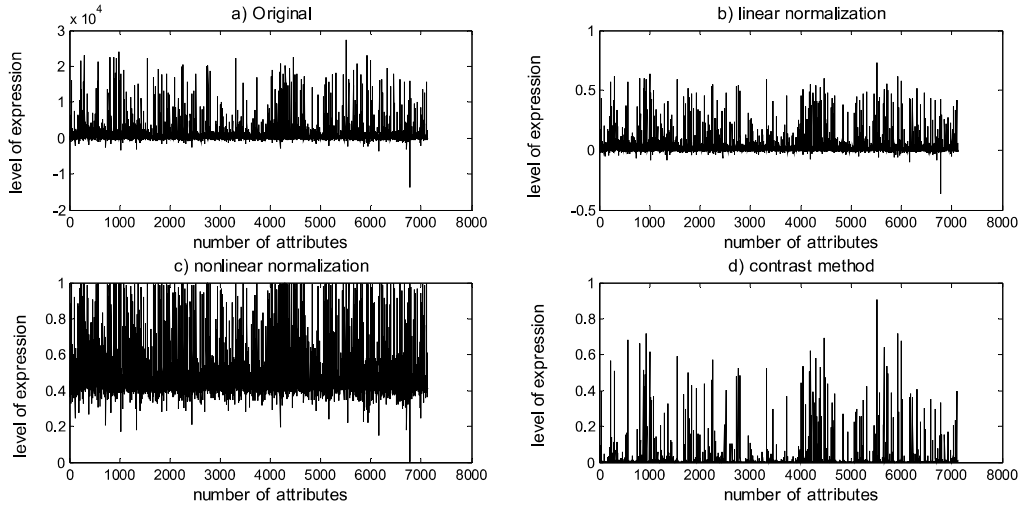
–calculate of normalizing coefficients by the formula:

$$x'_{ki} = 2^{\left(A_{ki} + \frac{\delta M_{ki}}{2}\right)}, \tag{16}$$

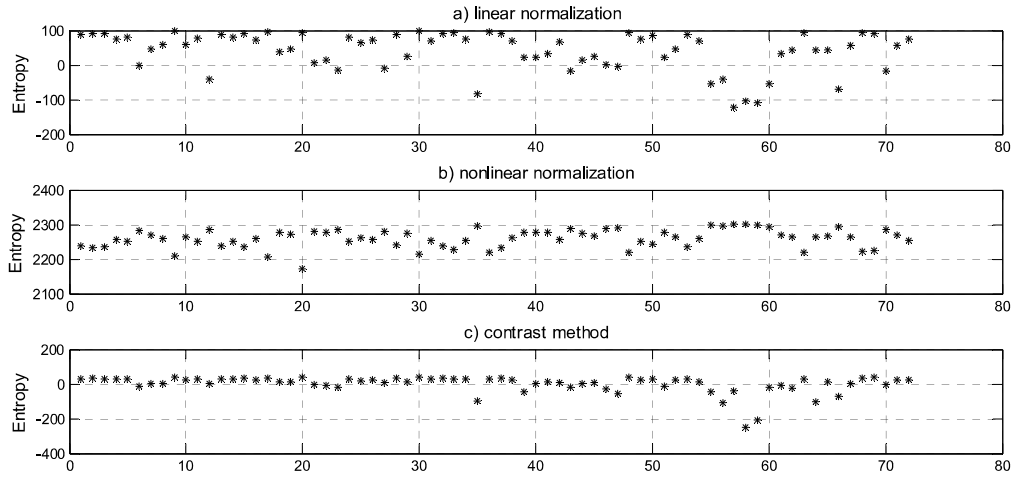
Database of leukemia patients (Golub et al, 1999) was used for research as an experimental database, represented itself array size 72×7131. Each row contains information about a level of gene expression of diseased cells of an individual human. The filtration of signals using the wavelet has been performing at the first step [8]. Necessity of the given step definice by high level of noise component in original data, arising on the stage of performing experiment and reading information from DNA microarray. As a result of wavelet signal processing it has been processing high-frequency component of the signal order to minimize the noise component. Hereinafter, each vector has been normalizing in accordance with above described algorithms. In each case for each normalized vectors was calculated the Shannon entropy in accordance with the quation (4). Plots of filtered signal of one investigated objects (Fig. 1a) and normalized signal linear normalization algorithm (Fig. 1c), nonlinear normalization (Fig. 1c) and contrast method (Fig. 1d) are showed on figure 1. Plots of entropy distribution of normalized signals depending on the method of normalization are showed in figure 2.

3. Discussion

Analysis of the graphs shown in Figures 1 and 2 allows to make a conclusion that from three methods of normalization the contrast method is more qualitative. The normalized signal shown in Figure 1d, has the least noise component than other signals. At the same time, as shown in Figure 2c, entropy of signals of investigated objects using the method of contrasts is minimal, that confirmed the assumption about effectiveness using the entropy criterion for evaluating of signal-to-noise ratio.



Picture 1 – Graphs of signals: a) original signal; b) normalized by linear normalization method; c) normalized by non-linear normalization method; d) normalized by contrast method.



Picture 2 – Graphic distribution of entropy of normalized signals investigated objects by using the method: a) linear normalization b) nonlinear normalization c) contrasts.

 ISSN 1562-9945

Effectiveness of using of the contrast method in this case can be explained by character of data distribution in investigated signal, which supposes an existence of linear regression dependence between analyzed characteristics. The method of nonlinear normalization gives the worst quality among using normalization methods. This can be explained by linear distribution of initial data that makes more efficient use of the linear normalization method. This conclusion confirmed by the entropy distribution graphs at using corresponding methods. Value of entropy of the signal in Fig. 2c is considerably higher than in Fig. 2a. It testifies lower value of signal-to-noise ratio in case of use of the non-linear normalization method, that confirmed by visual observations. The graph of normalized signal in Fig. 1c is noisier compared with the graph shown in Fig. 1c.

4. Conclusion

The problem of DNA microarray analysis nowadays is one of the actual problems of modern bioinformatics. Its solution will allow to forecast development of many genetic and epidemiology diseases with purpose to treat them on time. Preprocessing data is an essential step in the process of data analysis, which predetermines the quality of obtained final information. Data normalization is one of the major stages of preprocessing data. Experiments showed that qualitatively executed normalization allows to transform data to range convenient for further processing, while increasing the value of signal-to-noise ratio. An efficiency of using criteria Shannon entropy for evaluating quality of data normalization is showen in the article. Improving of quality of data normalization is accompanied by simultaneous entropy decreasing, which means increasing of informativeness of secondary signal. In future the author is planing to develop a system of clustering objects DNA microarrays, where presented research will be used for selecting optimum set of data preprocessing methods.

REFERENCES

1. N. Morrison, D. C. Hoyle. Concepts and Methods for Normalizing Microarray Data // Kluwer academic publishers.- 2003.- P. 76-90.
2. B. M. Bolstad, R. A. Irizarry, M. Astrand and T. P. Speed. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias // Bioinformatics.- 2003.- V. 19.- P. 185-193.

3. M. Astrand. Contrast Normalization of Oligonucleotide Arrays // Journal of Computational Biology.- 2003. V. 10(1).- P. 95-102.
4. P. Shambadal. Development and application of entropy. - M.: Science, 1967. - 280 p.
5. N. Martin, J. England. The mathematical theory of entropy.M.: World, 1988. - 350 p.
6. C.E.A. Shannon. Mathematical Theory of Communication // Bell System Technical Journal. - 1948. - V. 27. - P. 379-423, 623-656.
7. I.E.Irodov. Quantum physics. Basic laws // M.: Laboratory of Basic Knowledge, 2002. – 272 p.
8. S.A.Babichev, N.I.Babenko, A.A.Didyk, V.I.Litvinenko, A.A.Fefelov, S.V.Shkurdoda. Filtration of the chromatogram based on wavelet transform with use of entropy criterion // System technologies. - 2010. - N.6 (71). - S. 17-22.

 ISSN 1562-9945