AN IMPROVED K-NEAREST NEIGHBORS ALGORITHM FOR THE ANALYSIS OF TWO-COLOR DNA MICROARRAY DATA WITH SPOT QUALITY FACTORS

A. Svidrytski[1], M. Yatskou, V. Apanasovich
Belarusian State University
Minsk, BELARUS
e-mail: [1]svidrytski@gmail.com

Abstract

Algorithms for the classification of gene expression data require the continuous improvement of their efficiency. This paper presents a modification of a k-nearest neighbors algorithm which increases an efficiency of classification including quality parameter of each microarray spot. The efficiency of classification is achieved by recalculation of distances between classified and classifying objects. We also introduce an enhanced microarray data simulation model that includes spot quality parameters.

1 Introduction

Microarrays are one of the newest instruments of biology and medicine [1]. Their advantage caused by possibility to conduct an enormous number of specific reactions and interactions of biopolymer molecules simultaneously.

The first step in the entire microarray analysis is DNA microarray image processing, using such tools as GenePix or MAIA (MicroArray Image Analysis) [3]. Every mistake made on this step can further influence the final results significantly.

As a rule, after retrieving data, analysis of microarray gene expression includes step that removes objects with low quality [3]. In this paper there is presented an improved kNN algorithm, using quality parameter as a weight factor, which makes it possible to increase efficiency of a microarray analysis taking into account spots with lower quality. There is also presented algorithm for microarray simulation adapted for inclusion of quality parameter. This microarray model was applied to evaluate efficiency of the modified algorithm in comparison with original one.

2 DNA microarray model with a quality parameter

2.1 MAIA and a quality parameter

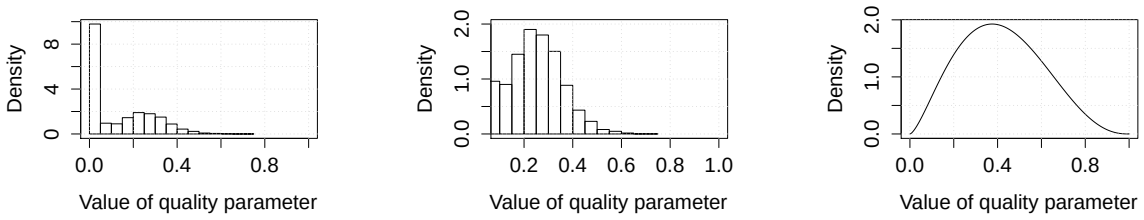
Microarray image analysis software like GenePix or MAIA provides an integrative estimate of spot quality in range from 0 to 1 after image processing [3]. The quality value can be used as a weight factor for classification analysis.

2.2 DNA microarray simulation model with a quality parameter

Simulation model of microarray gene expression values must be as similar to real data as possible. This is achieved by simulation of physical phenomena in the model. Dembele [2] proposed the universal microarray simulation model that is most advanced up to date. Thus we took this model for adding quality factor.

A distribution of the quality parameter was obtained resting on the microarray data in a whole-genome microarray experiment assessing well-characterized transcriptional modifications induced by the transcription regulator SNAI1 [3].

Histogram of the total spot quality parameters is shown in figure 1*a*. Such low-quality spots appear mainly due to poor microarray experiment conduction. Filtering objects with very low quality (from 0 to 0.1) resembles shape of beta-distribution with mode about 0.25 (figure 1*b*). Data of better quality would have mode higher than 0.25. In this work we took mode for the distribution equal 0.375. Empirical choice of the parameters for beta distribution equals to $\alpha_1 = 2.5$ and $\alpha_2 = 3.5$. Plot of density function with these parameters is presented in the figure 1*c*.



a *b* *c*

Figure 1: Histograms and Probability Density Function of the spot quality parameter: *a)* histogram of microarray data taken from [3] *b)* histogram of microarray data after filtering objects with qualities between 0 and 0.1 *c)* curve of beta-distribution with fitted parameters; it approximately resembles the shape of the histogram with shifted mode.

The next step is to bind the value of the quality parameter to value of gene expression that was implemented adding a Gaussian noise. Variance of the distribution is the greater the worse object quality is. The variance for i-th spot was calculated according to expression $\sigma_i = 1 - q_i$.

3 kNN modification using a quality parameter

The idea for the modification of the original k-nearest neighbors method involves adjusting distances between objects of training and test samples. Adjustments weight the between-object distances so that if the lower quality of training object then it contributes less to classifying test objects. The illustration of this approach is shown in

figure 2.

Algorithm:

1. Initialize the number of nearest neighbors k, the volumes of training and test samples.
2. Find distances d_{ij} between all objects of the training and test samples.
3. Modify distances with some function f which depends on quality parameter q_i: $d_{ij}^w = f(d_{ij})$.
4. Identify k nearest neighbors for each test object.
5. Determine the class label for each test object.

A function f was chosen hyperbolic: $d_{ij}^w = f(d_{ij}) = d_{ij}/qi$. This function is chosen because when a quality of spot goes to 0 then d_{ij}^w diverges to infinity. d_{ij} can be also logarithmically transformed: $d_{ij}^w = f(d_{ij}) = d_{ij} \cdot (-\log_a q_i + 1)$.

4 Results and discussion

4.1 Description of numerical experiments

Numerical experiment included:

1. Generation of test sample consisted of 10000 spots, where 100 spots are up regulated and 100 spots are down regulated.
2. Generation of training sample consisted of 300 spots where 100 spots are up regulated and 100 spots are down regulated. Expression values of neutral spots were distributed between -.1 and .1.
3. Receiving the number of correctly classified spots with classic and modified kNN methods considering the same test and training samples.

The described experiment was repeated 150 times under certain conditions. The changing conditions involved variation of the lowest value of spot quality for both training and test samples what reflected general microarray quality.

4.2 Results

The following formula was used to compare weighted and basic algorithms:

$$efficiency = 100\% \cdot \frac{T^{weighted}}{T^{basic}},$$

where $T^{weighted}$ and T^{basic} are the numbers of correctly classified objects by improved and basic algorithms respectively. Figure 3 represents heatmap, where the more green color reflects the more efficient improved algorithm is. Each cell contains average value of 150 efficiency values obtained for the certain lowest value of quality parameter.

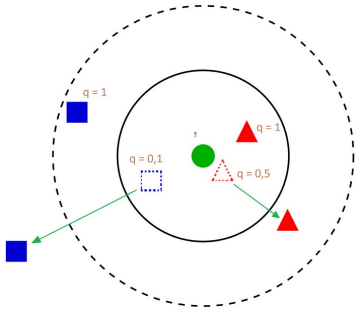


Figure 2: This picture illustrates how the method of distances modification works. A triangle with quality 0.5 is moved to a certain distance from the object for classification. A square which has lower quality 0.1 is moved even further. Squares and triangles that have quality 1 stay at the same place.

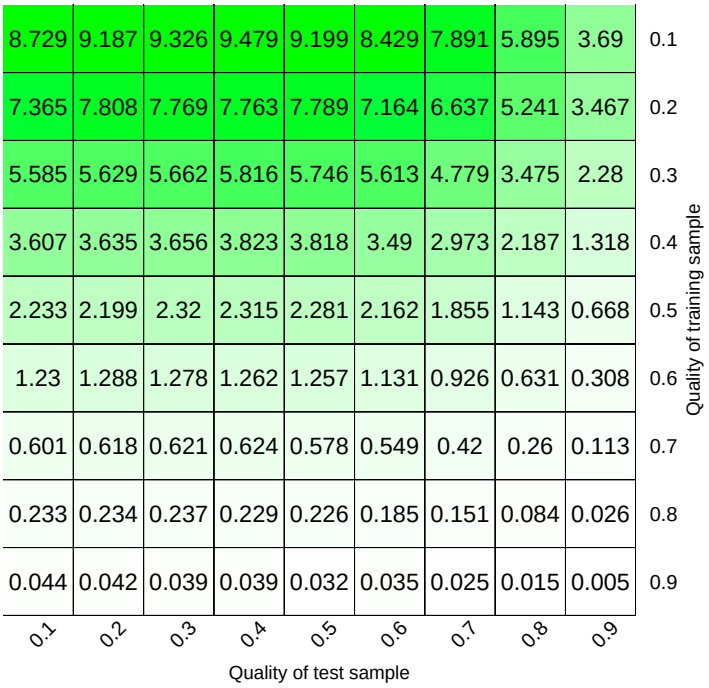


Figure 3: Heatmap of percent of efficiency of improved method in comparison with basic algorithm of kNN

5 Conclusion

This work presents a modified kNN algorithm that takes into account a spot quality parameter of a microarray. The algorithm works more efficiently than the classical one, when quality of training and test samples is worse. The idea of using the spot quality parameter can be embedded into more advanced algorithms of classification and cluster analysis.

References

[1] Zhang L. et al. (2015). Whole transcriptome microarrays identify long non-coding RNAs associated with cardiac hypertrophy. *Genom Data*. Vol. **5**, pp. 68–71.

[2] Dembele D.A. (2013). A Flexible Microarray Data Simulation Model. *Microarrays*. Vol. **2**, pp. 115–130.

[3] Yatskou M. et al. (2008). Advanced spot quality analysis in two-colour microarray experiments. *BMC Research Notes*. Vol. **1**(80).