

LINEAR PREDICTIVE CODING FOR ENHANCED MICROARRAY DATA CLUSTERING

Robert S. H. Istepanian, Ala Sungoor and Jean-Christophe Nebel

Mobile Information and Network Technologies Research Centre, Kingston University
Kingston-Upon-Thames, London, KT1 2EE, UK
r.istepanian@kingston.ac.uk, k0422190@kingston.ac.uk, j.nebel@kingston.ac.uk

ABSTRACT

Microarrays are powerful tools for simultaneous monitoring of the expression levels of large number of genes. Their analysis is usually achieved by using clustering techniques. In this paper, we present a new clustering method based on Linear Predictive Coding to provide enhanced microarray data analysis. In this approach, spectral analysis of microarray data is performed to classify samples according to their distortion values. The technique was validated for a standard data set. Comparative analysis of the results indicates that this method provides improved clustering accuracy compared to some conventional clustering techniques. Moreover, our classifier does not require any prior training procedure.

1. INTRODUCTION

It is well known that microarray technologies are one of the most powerful tools for extracting and interpreting simultaneous gene activities and relevant genomic information. In particular, analysis of microarray genetic data allows a better understanding of genetically based diseases such as diabetes, cardiovascular diseases and some forms of cancer. Although recently work has been reported on microarray imaging, signal processing and spot analysis techniques [1,2], gene classification and clustering is still an active field of research in this area. In general, gene expression analysis is based on statistical methods that are capable of detecting relevant genomic patterns that reflect individual genes in different regulatory states. The importance of this research area is reflected in the large number of works published in recent years [3].

Clustering analysis can be defined as a multivariate technique for data mining, the objective of which is to discover meaningful subgroups or objects. In microarray data analysis, clustering is concerned with the study of the associated gene expression matrix to group together either co-expressed genes or samples within similar gene expression profiles.

In recent years, various genomic signal processing methodologies were introduced for different microarray

application areas to support different clustering studies. These were used for detection, prediction, classification and in statistical modeling [4,5]. In particular, techniques based on Bayes vector quantization were applied to achieve better classification quality by minimization of the weighted summation of the mean squared error between a feature vector and its codeword structure [6-8]. Other classification methods based on spectral component analysis were also investigated. An Autoregressive technique was used to evaluate the potential regulatory relationship between genes with dominant spectral components in [9]. Also, in the work reported in [10], the expression profiles were decomposed into spectral component to correlate the profiles to obtain high accuracy expression values. However, to date no work has been reported for the application of advanced coding methods for enhanced microarray data clustering and gene expression analysis.

In this paper we present a new clustering framework based on the use of the Linear Predictive Coding for enhanced microarray data clustering. After providing an overview of the method, we present the application of the algorithm and validate our approach on a standard microarray data set.

2. PRINCIPLES OF LPC-BASED CLUSTERING

The Linear Predictive Coding (LPC) approach is based on estimation of spectral distortion measures that provide relative measures of gene expression changes. Therefore, enhanced classification of genes or samples into classes can be performed according to their distortion values.

Microarray data is first preprocessed and filtered using probabilistic estimators. Then gene expression data is transformed into distortion measures. The LPC algorithm is applied in order to build a predictive model for those data. Subsequently, LPC coefficients are converted into Line Spectral Frequency (LSF) coefficients to increase their spectral robustness. Finally, Vector Quantization (VQ) is applied to cluster the resultant data into the relevant classes. This clustering process is summarized in Figure 1.

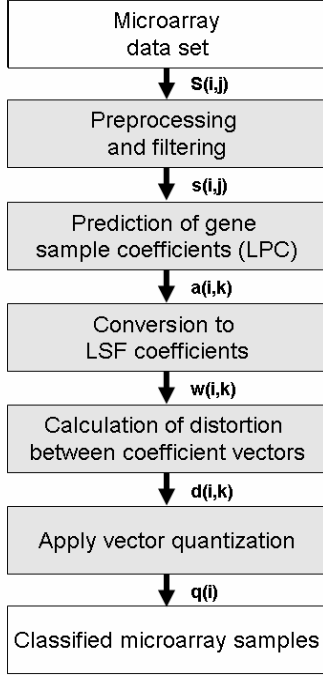


Figure 1. LPC-based microarray clustering method

2.1. LPC coefficient prediction of microarray data set

Features of expression sampling rate in the microarray data set format produce a limited range of frequencies. In addition, there are correlations both between genes and between samples. LPC in general is a coding method which is suitable to deal with data with such properties which are common in speech and imaging data. The basic idea behind LPC analysis is that each expression sample is approximated as a combination of past samples [11]. Equation (1) defines the LPC principle where the value of the present output, $s(n)$, can be predicted approximately by a linear combination of p past samples; p is called the order of LPC.

$$s(n) = \sum_{j=1}^p a_j s(n-j) \quad (1)$$

The goal of the LPC analysis is to find the best prediction coefficients a_j so that the predicted sample is a good approximation of the original sample. This optimization process is performed by minimizing the energy of the prediction error. This involves choosing a_j to minimize the mean energy, E , in the error signal over a frame or window of data set:

$$E = \left\{ \sum_{n=-\infty}^{\infty} \left[s(n) - \sum_{j=1}^p a_j s(n-j) \right]^2 \right\} \quad (2)$$

The values of a_j that minimize E are found by setting all derivatives dE / da_j equal to zero. It is expressed by:

$$\sum_{i=1}^p a_i E(s_{n-i} s_{n-j}) = E(s_n s_{n-j}) \quad (3)$$

To solve equation (3), $E(s_{n-i} s_{n-j})$ needs to be estimated for $i, j \in \{1, \dots, p\}$. The autocorrelation and covariance methods are two of the most common and efficient linear predictive spectral estimation techniques. Their main difference lies in the placement of the analysis window. Since the covariance method windows the error signal instead of the original signal, it has a highest accuracy.

The energy E of the windowed error signal is:

$$E = \sum_{n=-\infty}^{\infty} e^2(n)w(n) = \sum_{n=-\infty}^{\infty} \left[s(n) - \sum_{k=1}^m a_k s(n-k) \right]^2 w(n) \quad (4)$$

where the error is minimized over a finite interval of size N as defined by the rectangular window function $w(n)$.

After reducing and differentiating equation (4) with respect to a_k , we obtain:

$$\sum_{n=0}^{N-1} s(n-i)s(n) = \sum_{k=1}^M a_k \sum_{n=0}^{N-1} s(n-k)s(n-i) \quad (5)$$

However, the process of direct quantization of the LPC coefficients a_j is not advisable. The issue is that small changes due to the quantization error could result in the internal digital filter pole becoming unstable and producing large spectral errors. Thus, other superior parametric representations have been formulated to replace the LPC coefficients a_j [12]. They include log area ratios [13], arcsine reflection coefficients [14] and line spectral frequency (LSF) [15]. We chose the LSF representation because it has been shown to be a particularly efficient for scalar quantization of LPC information: it also does not distort the spectrum, vary smoothly in time and offers a better coding in relation to spectral peaks. These LSF coefficients are used subsequently to determine distortion between samples.

2.2. Predictive vector quantization

The use of vector quantization in this application is twofold: firstly we want to capture meaningful classes in the data, represented by their centers, and secondly, we want to make our subsequent classification decisions more robust to noise in the data.

The principle of Vector Quantization (VQ) is to map L -dimensional input vectors into a set of M vectors $q(i)$ (with $L > M$). These vectors are called code vectors or codewords and C is called the codebook. The average quantization error between input source and their reproduction codeword is called the distortion of the vector quantizer. The major concern for a vector quantizer codebook design is the trade-off between distortion and rate. Once the number of quantization levels is defined, the rate is set. Then the focus is on data quantization as a means of removing noise from data. The centers of the groups of data corresponding to different quantization levels should be selected such that distortion is minimized.

In this work, we use a ‘nearest neighbor’ vector quantizer, i.e. a vector z is mapped to a code vector q_m if

$$d(z, q_m) = \arg \min_i (d(z, q_i)) \quad (6)$$

where d is a suitable distortion measure.

We use the gain-normalized log spectral distortion since it is widely accepted as a quality measure of coded speech spectra. It evaluates the similarity of two auto-regressive envelopes. There are expressed in the frequency domain by the following equation [16]:

$$d(z, q_i) = \int_{-p}^{+p} (\log P_z(\mathbf{w}) - \log P_{q_i}(\mathbf{w}))^2 \frac{d\mathbf{w}}{2p} \quad (7)$$

where $P(\mathbf{w})$ is the auto-regressive envelope that is defined as:

$$P(\mathbf{w}) = \frac{1}{|1 + \sum_{k=1}^p a_k e^{-jwk}|^2} \quad (8)$$

The design of codebooks is usually accomplished by an iterative algorithm called the Lloyd algorithm. This algorithm generates a set of representative vectors of the source data and optimizes the codebook using the distortion measure d [11].

Finally, once the codebook has been defined, LSF coefficient vectors of s are extracted, compared to all codewords of C and mapped to a single codeword.

3. RESULTS AND DISCUSSION

In order to validate the cognate LPC and VQ approach, a MATLAB® implementation was used for the classification of acute leukemia patients using a standard microarray data set defined by Golub et al. [17]. This data set was selected as it is now considered as a benchmark for microarray data classification. Acute leukemia can be divided into two classes, Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). Correct diagnosis of a patient is critical since those two diseases require different treatments. Golub et al. used data from two different sets: a 38 sample set composed of 11 AML and 27 ALL patients and a 34 sample set composed of 14 AML and 20 ALL patients. They used those two sets as training and test sets respectively. Since the LPC method does not rely on any form of training, we treated both sets as test sets. Moreover, we considered a combined set made of the two sets to investigate the effect of the size of the set on the performance of our technique.

We applied a statistical normalization method [18] to the microarray data. Then we selected the 50 genes with the highest expression values as shown in Golub et al [17]. Next we processed these data using a 10th order LPC algorithm. The predicted coefficients were then converted to LSF. Finally, classification was achieved by performing vector quantization using 2 codewords.

Figure 2 shows how the combined sample set clusters: samples are plotted according to their distortion distances to the two classes. A sample can then be assigned to the nearest class.

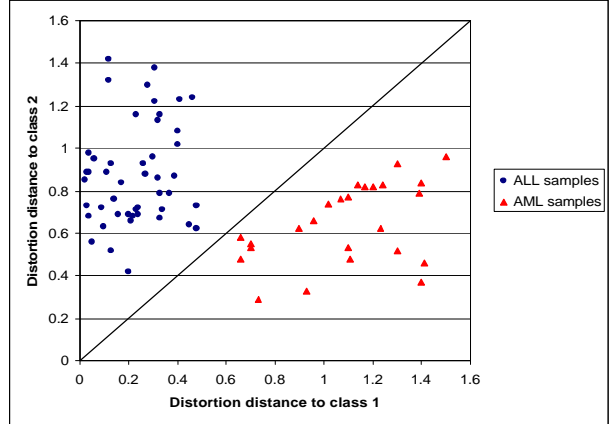


Figure 2. Clusters for the combined set of 72 samples.

The total computation time using a Pentium PC for the whole dataset was 1.8 seconds. The result of the method presented here is summarized in the last column of Table 1. It is clear from these results that the LPC method produces superior performance for the first set (i.e. Golub’s training set) since classifications were accurate for all 28 samples of the set. The second set (i.e. Golub’s test set) is more challenging because it includes a much broader range of samples coming from different sources. The LPC technique performed well since 30 samples were accurately classified while only 4 were incorrect. However, these 4 samples are particularly difficult to classify since they were among the 5 that Golub could not classify.

Finally, 100% classification with the combined set is achieved as shown in Figure 2. The increase in the number of samples allows better definition of the two classes. Therefore, it improves the discrimination power of the classifier and permits the correct classification of the most difficult samples. Table 1 summarizes the performances of existing methods developed by 5 other teams including Golub’s.

Table 1: Comparative performance of LPC clustering with existing methods

| | Golub [17] | Nguyen [19] | Yeung [20] | Tan [21] | Liu [22] | LPC |
|--------------|------------|-------------|------------|----------|----------|-----|
| Method | T-test | PLSLD | Bayesian | TPCR | KPCA | LPC |
| Supervised | Yes | Yes | No | No | Yes | No |
| Training set | 38 | 38 | / | / | 38 | 38 |
| Correct | 36 | 38 | / | / | 38 | 38 |
| Incorrect | 0 | 0 | / | / | 0 | 0 |
| Uncertain | 2 | 0 | / | / | 0 | 0 |
| Test set | 34 | 34 | / | / | 34 | 34 |
| Correct | 29 | 33 | / | / | 33 | 30 |
| Incorrect | 0 | 1 | / | / | 1 | 4 |
| Uncertain | 5 | 0 | / | / | 0 | 0 |
| Whole set | / | / | 72 | 72 | / | 72 |
| Correct | / | / | 70 | 71 | / | 72 |
| Incorrect | / | / | 2 | 1 | / | 0 |
| Uncertain | / | / | 0 | 0 | / | 0 |

4. CONCLUSION

In this paper we presented a new clustering method based on Linear Predictive Coding. Spectral analysis of microarray data was performed to classify samples according to their distortion values. This technique combined with a Vector Quantization approach was validated using a standard microarray data set. Comparative analysis of the results indicates that the LPC method provides improved clustering accuracy compared to some conventional clustering methods. Moreover, this classifier does not require any form of training. The only limitation of this method is its performance relies on the size of the set to produce robust classification. Further work is currently underway to compare this microarray clustering approach with other digital signal processing methodologies.

5. REFERENCES

- [1] R. Istepanian, "Microarray image processing: current status and future directions", *IEEE transactions on Nanobioscience*, vol. 2(4), pp. 173-175, Dec. 2003.
- [2] X. H. Wang, R. Istepanian and T. Geake, "Error Control Coding in Microarray Data Analysis", in *Proc. Int. Workshop on Genomic Signal Processing and Statistics*, Baltimore, USA, May 2004.
- [3] T. D. Pham, C. Wells and D. Crane, "Analysis of microarray gene expression data", *Current Bioinformatics*, vol. 1, pp. 37-53, 2006
- [4] W. Zhang and I. Shmulevich, Eds., *Computational and Statistical Approaches to Genomics*: Kluwer Academic Publishers, Boston, 2002.
- [5] T. Kohonen, *Self-Organization and Associative Memory*, Springer-Verlag, New-York, 1988.
- [6] K. L. Oehler and R. M. Gray, "Combining image compression and classification using vector quantization", *IEEE Trans. Pattern analysis and machine intelligence*, vol. 17(5), pp. 461-473, May 1995.
- [7] K. O. Perlmutter, S. M. Permuter, R. M. Gray, R. A. Olshen, and K. L. Oehler, "Bayes risk weight vector quantization with posterior estimation for image compression and classification", *IEEE Trans. image Processing*, vol. 5(2), pp. 347-360, 1996.
- [8] J. Li and H. Zha, "Simultaneous classification and feature clustering using discriminant vector quantization with application to microarray data analysis", In *Proc. IEEE computer society Bioinformatics Conf.*, Stanford, USA, Aug. 2002.
- [9] L. K. Yeung, L. K. Szeto, A-C. Liew and H. Yan, "Dominant spectral component analysis for transcriptional regulations using microarray time-series data", *Bioinformatics*, vol. 20, pp. 742-749, 2004.
- [10] P. Spellman, G. Sherlock, M. Zhang et al., "Comprehensive identification of cell cycle regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization", *Mol. Biol. Cell*, vol. 9, pp. 3273-3297, 1998.
- [11] T. Quatieri, *Discrete Time speech signal processing*, Prentice Hall, 2002.
- [12] K. K. Paliwal and W. B. Kleijn, "LPC Quantization", in *Advances in Speech Coding*, pp. 433-466, Norwell, MA, USA, Kluwer, 1991.
- [13] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems", in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp. 309-321, 1975.
- [14] A. H. Gray and J. D. Markel, "Quantization and bit allocation in speech processing", in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp.459-573, 1976.
- [15] F. Itakura, "Line spectrum representation of linear predictive coefficients", *J. Acoust. Soc. Am.*, vol. 57 suppl. 1, pp.35, 1975.
- [16] D. O'Shaughnessy, *Speech Communication: Human and Machine*, New York, IEEE Press, 2000.
- [17] T. R. Golub, D. K. Slonim et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", *Science*, vol. 286, pp. 531-538, 1999.
- [18] D. J. Balding, M. Bishop, C. Cannings, "Handbook of Statistical Genetics", John Wileyand Sons, 2001
- [19] D. V. Nguyen and D. M. Rocke, "Tumor classification by partial least squares using microarray gene expression data", *Bioinformatics*, vol. 18(1), pp. 39-50, 2002.
- [20] K. Y. Yeung, R. E. Bumgarner and A. E. Raftery, "Bayesian Model Averaging: development of an improved multi-class, gene selection and classification tool for microarray data", *Bioinformatics*, vol. 21(10), pp. 2394-2402, 2005.
- [21] Y. Tan, L. Shi, W. Tong and C. Wang, "Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data", *Nucleic Acids Research*, vol. 33(1), pp. 56-65, 2005.
- [22] Z. Liu, D. Chen and H. Bensmail, "Gene Expression Data Classification With Kernel Principal Component Analysis", *J. Biomed. Biotechnol.*, vol. 2, pp. 155-159, 2005.