ESTIMATING THE NUMBER OF CLUSTERS IN MICROARRAY DATA SETS BASED ON AN INFORMATION THEORETIC CRITERION

Daniel Nicorici, Jaakko Astola, Olli Yli-Harja

Institute of Signal Processing, Tampere University of Technology
P.O. Box 553, FIN-33101 Tampere, Finland
E-mails: Daniel.Nicorici@tut.fi, Jaakko.Astola@tut.fi, Olli.YliHarja@tut.fi

ABSTRACT

This study focuses on an information theoretic approach for estimating the number of clusters K, in microarray data sets. We present an automatic method for estimating K, based on a particular version of the Normalized Maximum Likelihood (NML) model. The strength of the Minimum Description Length (MDL) methods, such as the NML model, in statistical inference is to find the model structure which, in this particular clustering problem, amounts to find the best number of clusters and the best cluster structure for the data. The models are compared using the NML code length. The study introduces a new method for computing the code length of the encoded clustering vector for the data samples, based on the NML model. Experiments with publicly available microarray data sets demonstrate the ability of the new method to find the biologically meaningful clusters.

Keywords: number of clusters, microarray data, minimum description length, normalized maximum likelihood.

1. INTRODUCTION

Unsupervised analysis of microarray data sets is of particular significance for gene expression data analysis. The accurate estimation of the number of clusters K, is particularly important for gene expression data analysis because most of the existing clustering procedures request K as input. Recent proposals exist for estimating the number of clusters K in a data set based on cluster stability [1, 2], resampling methods [3], Bayesian model-based method [4], Gap statistics [5], and Minimum Description Length (MDL) principle [6, 7, 8].

Recently, a modern and more efficient form of MDL principle based on the Normalized Maximum Likelihood (NML) model [9], which effects a universal sufficient statistics decomposition to separate noise from the learnable information in data, has been found [10]. The NML model and the MDL principle [11] have been used successfully before in genomic signal processing for classification and feature selection problems [10], simultaneous gene clustering [12], fast iterative gene clustering [7], compression of DNA sequences [13], and for finding large domains of similarly expressed genes [14].

We propose a new method for estimating the number of clusters that makes use of the NML model for the K–Means clustering algorithm with the Euclidean distance. We restrict our investigation to the agglomerative K–Means clustering algorithms mainly because this class of clustering methods is very popular in microarray data analysis [15]. According to the MDL principle [11], the model is selected based on its fitting performance but also penalizing a too high complexity of the model. The MDL principle has been widely used in statistical inference in an asymptotically justified form [10]. The MDL principle is used in this study for evaluating the model for clustering, based on the idea that a good clustering is such that one can encode the clustering together with the data such that the resulting code length is minimized [8]. The new method uses the NML model for data clustering, which is a refinement of the method introduced by Kontkanen et al. [8], with an improved method for computing the code length of the clustering vector.

We study in the first part of the paper the modeling of the data with clusters using the MDL principle and the NML model. We then proceed by introducing the new method that contains the NML model and the K–Means clustering algorithm. Finally, we apply the new method to estimate the number of clusters for publicly available microarray data sets and evaluate its effectiveness in discovering biologically meaningful clusters.

2. MODEL BASED DATA CLUSTERING

We present here the MDL principle for data clustering and the NML model approach for (i) encoding an q-ary sequence, for (ii) encoding the cluster data, and for (iii) encoding the clustering vector.

2.1. MDL principle

The MDL principle [11] considers the description length of the data D and the model $\mathcal{M}$ as follows

$$\mathcal{L}(\mathcal{M}, D) = \mathcal{L}(\mathcal{M}) + \mathcal{L}(D|\mathcal{M}), \quad (1)$$

where $\mathcal{L}(D)$ is the length of the model's description and $\mathcal{L}(D|\mathcal{M})$ is the length of data's description, where the data is described using the model $\mathcal{M}$. According to the MDL principle the best model for the data is the model with the shortest length of the total description $\mathcal{L}(\mathcal{M}, D)$.

2.2. The NML model for q-ary sequences

Let us consider a sequence $\mathbf{z}^n = [z_1, \ldots, z_n]$ where $z_i \in \{1, \ldots, q\}$ for $i = 1, \ldots, n$. The NML approach is used for encoding the q-ary sequence $\mathbf{z}^n$ by postulating a simple parametric model $P(\mathbf{z}^n; \Theta(\mathbf{z}^n))$. The sequence $\mathbf{z}^n$ contains in general h_i values of i. For a q-ary sequence $\mathbf{z}^n$, the ML estimate of $\Theta = \{\theta_1, \ldots, \theta_q\}$ is $\hat{\Theta}(\mathbf{z}^n) = \{\hat{\theta}_1(\mathbf{z}^n), \ldots, \hat{\theta}_q(\mathbf{z}^n)\}$, where $\hat{\theta}_i(\mathbf{z}^n) = \frac{h_i}{n}$. The probability of the $\mathbf{z}^n$ becomes $P(\mathbf{z}^n; \hat{\Theta}(\mathbf{z}^n)) = \prod_{i=1}^{q} (\frac{h_i}{n})^{h_i}$. The NML model is known to be a solution of two minmax problems, which gives it strong optimality properties [9]. The normalized maximum likelihood is

$$\hat{P}(\mathbf{z}^n; \hat{\Theta}(\mathbf{z}^n)) = \frac{\prod_{i=1}^{q} \left(\frac{h_i}{n}\right)^{h_i}}{R_q^n}, \quad (2)$$

where

$$R_q^n = \sum_{\mathbf{y}^n \in \mathcal{Z}^n} P(\mathbf{y}^n; \hat{\Theta}(\mathbf{y}^n)), \quad (3)$$

is the normalizing factor for the probability $P(\mathbf{z}^n; \hat{\Theta}(\mathbf{z}^n))$ and $\mathcal{Z}^n$ is chosen to include all the possible q-ary sequences $\mathbf{y}^n$. Thus, one has

$$R_q^n = \sum_{h_1 + \ldots + h_q = n} \frac{n!}{h_1! \ldots h_q!} \prod_{i=1}^{q} \left(\frac{h_i}{n}\right)^{h_i}, \quad (4)$$

for $h_1, \ldots, h_q \geq 0$. Kontkanen et al. [8] introduced an efficient way of computing the R_q^n, as following

$$R_q^n = \sum_{r_1 + r_2 = n} \frac{n!}{r_1! r_2!} \left(\frac{r_1}{n}\right)^{r_1} \left(\frac{r_2}{n}\right)^{r_2} \cdot R_{q^*}^{r_1} \cdot R_{q-q^*}^{r_2}, \quad (5)$$

where $R_q^0 = 1$, $R_1^n = 1$, $q^* = 1$, and $r_1, r_2 \geq 0$. The values for R_q^n can be pre-computed and tabulated to speed up the computations. The code length in bits for the encoded q-ary sequence $\mathbf{z}^n$ using the NML model is

$$\begin{aligned} \mathcal{L}_{NML}(\mathbf{z}^n) &= -\log_2 \hat{P}(\mathbf{z}^n; \hat{\Theta}(\mathbf{z}^n)) \\ &= \log_2 R_q^n - \log_2 \prod_{i=1}^{q} \left(\frac{h_i}{n}\right)^{h_i}. \end{aligned} \quad (6)$$

2.3. The NML models for data clustering

In this study, we use a refinement of the model-based approach introduced by Kontkanen et al. [8] for estimating the number of clusters and finding the cluster-structure of the data. The approach is based on the idea that a good clustering is such that one can encode the clustering together with the data, using the NML model, so that the resulting code length is minimized [8]. The models, representing the given cluster-structures of the data, are compared using the NML code length. The problems of finding the number of clusters and the cluster structure are solved simultaneously by choosing the best model, which gives the shortest NML code length for the data.

Let us consider a data set $\mathbf{x}^n = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ consisting of n column vectors, where $\mathbf{x}_i = [x_{i1}, \ldots, x_{im}]^T$, for $i = 1, \ldots, n$. The x_{ij} are quantized to q levels, such that $x_{ij} \in \{1, \ldots, q\}$ for $j = 1, \ldots, m$. Clustering the data set $\mathbf{x}^n$ is defined as a partitioning of the data into disjoints subsets whose union forms the data set. Each clustering is represented by the clustering vector $\mathbf{y}^n = [y_1, \ldots, y_n]$ where $y_i \in \{1, \ldots, K\}$ and $y_i = k$ if and only if $\mathbf{x}_i$ belongs to the cluster k. The number of clusters is denoted by K where $K = 1, \ldots, n$.

The NML code length, in bits, of the encoded data $\mathbf{x}^n$ together with the clustering vector $\mathbf{y}^n$ considering the model $\mathcal{M}_K$ [8], where K is the number of clusters, is

$$\begin{aligned} \mathcal{L}(\mathbf{x}^n, \mathbf{y}^n | \mathcal{M}_K) &= \log_2 R_K^n - \log_2 \prod_{i=1}^{K} \left(\frac{h_i}{n}\right)^{h_i} \\ &+ m \cdot \log_2 \prod_{i=1}^{K} R_q^{h_i} - \log_2 \prod_{i=1}^{m} \prod_{k=1}^{K} \prod_{v=1}^{q} \left(\frac{f_{ikv}}{h_k}\right)^{f_{ikv}}, \end{aligned} \quad (7)$$

where h_k is the number of times y_i such that $y_i = k$ [8], f_{ikv} is the number of data points x_{iu} such that $\mathbf{x}_i$ belongs to clusters k and $x_{iu} = v$ [8], and R_K^n, $R_q^{h_i}$ are the normalization factors computed using (5). The first two terms in (7) give the cost in bits of encoding the clustering vector $\mathbf{y}^n$ and the last two terms are needed to encode the data $\mathbf{x}^n$ given the clustering vector $\mathbf{y}^n$.

The estimated number of clusters for a data set $\mathbf{x}^n$ is given by the value of K, where $K = 1, \ldots, n$, for which the NML code length $\mathcal{L}(\mathbf{x}^n, \mathbf{y}^n | \mathcal{M}_K)$ is minimum.

2.4. A new NML model for encoding the clustering vector

In the approach of Kontkanen et al. [8] the clustering vector $\mathbf{y}^n$ is encoded considering all possible q-ary sequences of length n, as in (6). We introduce a novel way of computing the code length of the encoded clustering vector $\mathbf{y}^n$ by taking into account more faithfully the number of clustering vectors. This allows us to encode more efficiently the

clustering vector using the NML model and to discriminate better between different models. The normalization factor for encoding the clustering vector $\mathbf{y}^n$ with K clusters, using the NML model, is

$$S_K^n = \sum_{t^n \in \mathcal{F}^n} P(t^n; \hat{\Theta}(\mathbf{t}^n)), \tag{8}$$

where $\mathcal{F}^n$ is chosen to include all possible and unique clustering vectors $\mathbf{t}^n$ with K clusters. Two clustering vectors (partitions) are considered unique if they do not agree, i.e. RAND index $\neq 1$. The RAND index [16] ranges between 0 and 1 and it is a measure of agreement between alternative data partitions (data clusters).

For instance, one has for $n = 4$ and $K = 3$, the following space $\mathcal{F}^4$ of 6 unique clustering vectors

$$\begin{array}{cccc} 1 & 2 & 3 & 3 \\ 1 & 3 & 2 & 3 \\ 3 & 1 & 2 & 3 \\ 1 & 3 & 3 & 2 \\ 3 & 1 & 3 & 2 \\ 3 & 3 & 1 & 2 \end{array}$$

instead of 36 possible clustering vectors which are not all unique between them.

One can notice that the normalization factor in (8) for encoding the clustering vector $\mathbf{y}^n$ with K unique clusters becomes

$$S_K^n = \frac{R_K^n(r_1, r_2 > 0)}{K!}, \tag{9}$$

where $R_K^n(r_1, r_2 > 0)$ is computed using (5) such that $r_1, r_2 > 0$. The code length, in bits, of the encoded clustering vector $\mathbf{y}^n$ with K clusters using the NML model is

$$\mathcal{L}_{NML}(\mathbf{y}^n) = \log_2 S_K^n - \log_2 \prod_{i=1}^{K} \left(\frac{h_i}{n}\right)^{h_i}. \tag{10}$$

For instance, for a clustering vector $\mathbf{y}^n$ such that $n = K$, one has from (10) that $\mathcal{L}_{NML}(\mathbf{y}^n) = 0$, which is what one would expect. In this case no additional information is needed for encoding the clustering vector beside the specified model order. It is enough for the decoder to know that $n = K$ in order to decode the clustering vector. When (6) is used, one has $\mathcal{L}_{NML}(\mathbf{y}^n) > 0$ and the code length of the encoded clustering vector is longer than in the previous case.

The new NML code length, in bits, of the encoded data $\mathbf{x}^n$ together with the clustering vector $\mathbf{y}^n$ considering the model $\mathcal{M}_K$, where K is the number of clusters, is

$$\mathcal{L}(\mathbf{x}^n, \mathbf{y}^n | \mathcal{M}_K) = \log_2 S_K^n - \log_2 \prod_{i=1}^{K} \left(\frac{h_i}{n}\right)^{h_i} + m \cdot \log_2 \prod_{i=1}^{K} R_q^{h_i} - \log_2 \prod_{i=1}^{m} \prod_{k=1}^{K} \prod_{v=1}^{q} \left(\frac{f_{ikv}}{h_k}\right)^{f_{ikv}}, \tag{11}$$

where S_K^n is computed using (9) and h_k, f_{ikv}, R_K^n, and $R_q^{h_i}$ are as in (7).

3. CLUSTER STRUCTURE INFERENCE

The new method for estimation of number of clusters and finding the cluster structure in a data set $\mathbf{x}^n$ proceeds as follows. For each $K = 1, \ldots, n$ one clustering vector $\mathbf{y}^n$ with the best K clusters are found using the K–Means clustering with the Euclidean distance. The clustering vector $\mathbf{y}^n$ with its corresponding K that gives the shortest NML code length $\mathcal{L}(\mathbf{x}^n, \mathbf{y}^n | \mathcal{M}_K)$, computed using (11), is considered the one which provides the best estimation for (i) the cluster structure, and (ii) the number of clusters.

4. EXPERIMENTAL RESULTS

We illustrate the estimation of number of clusters on several real publicly available microarray data sets. The microarray data sets: "Leukemia", "Novartis", "St. Jude", "Lung Cancer", "CNS tumors", and "Normal tissues" are described and preprocessed as in [3]. Further, we quantize independently each gene profile of each microarray data set to binary states ("express"/"not expressed") by applying the Lloyd algorithm [10].

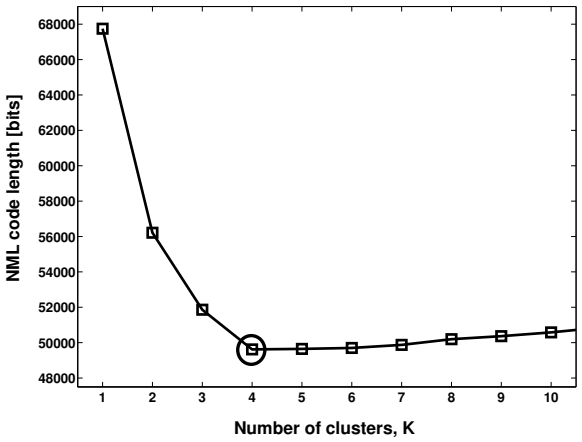

Fig. 1. The NML code length versus the number of clusters K, for Novartis microarray data set. The minimum of the NML code length is circled on the graph and it occurs for $K = 4$.

We apply the new method for estimating the number of clusters using the NML model for data clustering, as in (11), and the K–Means clustering, for the Novartis data set. The corresponding NML code lengths $\mathcal{L}(\mathbf{x}^n, \mathbf{y}^n | \mathcal{M}_K)$ for different number of clusters K, for the Novartis data set are as plotted in Figure 1. The minimum NML code length is

found for $K = 4$ clusters that corresponds to the biological knowledge.

Table 1. Estimated number of clusters using Consensus Clustering (CC) with Hierarchical Clustering (HC) and NML model with K–Means clustering, for microarray data sets

Data set	K_{true}	$\mathrm{CC}_{\mathrm{HC}}$	$\mathrm{NML}_{\mathrm{K-Means}}$
Leukemia	3	5	5
Novartis	4	**4**	**4**
St. Jude	6	5	**6**
Lung cancer	4+	5	9
CNS tumors	5	**5**	**5**
Normal tissues	13	7	8

Table 2. RAND index for Consensus Clustering (CC) with Hierarchical Clustering (HC), and NML model with K–Means clustering, for microarray data sets. In parentheses is RAND index corresponding to partitioning into K_{true} clusters when this differ from estimated K

Data set	$\mathrm{CC}_{\mathrm{HC}}$		$\mathrm{NML}_{\mathrm{K-Means}}$	
Leukemia	0.648	(1.000)	0.832	(1.000)
Novartis	0.921		0.980	
St. Jude	—	(0.948)	0.981	
Lung cancer	0.310	(0.280)	0.572	(0.745)
CNS tumors	0.549		0.835	
Normal tissues	0.457	(0.572)	0.869	(0.911)

Monti et al. [3] introduces the Consensus Clustering (CC) method for estimating the number of clusters for microarray data sets. They find that the CC with the Hierarchical Clustering (HC), notated as $\mathrm{CC}_{\mathrm{HC}}$, outperforms the CC with the Self-Organizing Map (SOM), the Gap with the HC, and the Gap with the SOM at estimating number number the clusters and agreement between the predicted and real clusters in microarray data sets. Thus we compare our method against the $\mathrm{CC}_{\mathrm{HC}}$ using the same microarray data sets and results reported in [3]. The comparison results between our newly introduced method, notated as $\mathrm{NML}_{\mathrm{K-Means}}$ and $\mathrm{CC}_{\mathrm{HC}}$, are shown in Table 1 and Table 2. We note that for the "St. Jude" data set in [3] the right number of clusters is found after "visual inspection of the consensus matrices". For the comparison we use only the values found automatically with both methods.

The new method, based on the NML and the K–Means clustering, outperforms the Consensus Clustering with the Hierarchical Clustering [3] at finding the true number of clusters in real microarray data sets (Table 1) and also in classification accuracy (Table 2).

5. CONCLUDING REMARKS

We have introduced a new method for estimating the number of clusters in microarray data sets based on the NML model for the K–Means clustering algorithm. The NML model is used for evaluating the goodness of clustering such that one can encode the clustering together with the data, where the data samples from the same cluster can be compressed well together and the resulting total code length is minimized. The problems of finding the number of clusters and the cluster structure are solved simultaneously by choosing the best model, which gives the shortest NML code length for the data. We have applied the new method to real microarray data sets in order to demonstrate its ability to find biologically meaningful clusters.

6. REFERENCES

[1] C.D. Giurcăneanu and I. Tăbuş, "Cluster structure inference based on clustering stability with applications to microarray data analysis," *Journal of Applied Signal Processing, Special Issue on Genomic Signal Processing*, vol. 1, no. 1, pp. 64–80, 2004.

[2] A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data," in *Pacific Symposium on Biocomputing*, 2002, pp. 6–17.

[3] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Journal of Machine Learning*, vol. 52, no. 1–2, pp. 91–118, July 2003.

[4] C. Fraley and A.E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," *The Computer Journal*, vol. 41, no. 8, pp. 578–588, 1998.

[5] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the Gap statistic," *Journal of the Royal Statistical Society*, vol. 63, no. 2, pp. 411–423, 2001.

[6] R.S. Wallace and T. Kanade, "Finding natural clusters having minimum description length," in *Proceedings of the 10th International Conference on Pattern Recognition*, June 1990, vol. 1, pp. 438–442.

[7] C.D. Giurcăneanu, I. Tăbuş, J. Ollila, and M. Vihinen, "Fast iterative gene clustering based on information theoretic criteria for selecting the cluster structure," *Journal of Computational Biology*, vol. 11, no. 4, pp. 660–682, 2004.

[8] P. Kontkanen et al., "An MDL framework for data clustering," in *Advances in Minimum Description Length: Theory and Applications*, P. Grünwald, I.J. Myung, and M. Pitt, Eds., pp. 323–353. MIT Press, Cambridge, 2005.

[9] J. Rissanen, "Strong optimality of the normalized ML models as universal codes and information in data," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1712–1717, 2001.

[10] I. Tăbuş, J. Rissanen, and J. Astola, "Classification and feature gene selection using the normalized maximum likelihood model for discrete regression," *Signal Processing, Special Issue on Genomic Signal Processing*, vol. 83, no. 4, pp. 713–727, 2003.

[11] J. Rissanen, "Modelling by the shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[12] R. Jörnsten and B. Yu, "Simultaneous gene clustering and subset selection for classification via MDL," *Bioinformatics*, vol. 19, no. 9, pp. 1100–1109, 2003.

[13] I. Tăbuş, G. Korodi, and J. Rissanen, "DNA sequence compression using the normalized maximum likelihood model for discrete regression," in *Proceedings of the IEEE Data Compression Conference (DCC'03)*, Snowbird, Utah, USA, 2003, pp. 253–262.

[14] D. Nicorici, O. Yli-Harja, and J. Astola, "An MDL method for finding large domains of similarly expressed genes," in *Workshop on Genomic Signal Processing and Statistics (GENSIPS'04)*, Baltimore, Maryland, USA, May 26–28, 2004.

[15] M. Sultan and et al., "Binary tree-structured vector quantization approach to clustering and visualizing microarray data," *Bioinformatics*, vol. 18, no. 1, pp. S111–S119, 2002.

[16] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985.