CLUSTERING CHANGE DETECTION USING NORMALIZED MAXIMUM LIKELIHOOD CODING

So Hirai[1], *Kenji Yamanishi*[2]

[1]Graduate School of Information Science and Engineering, The University of Tokyo,
7-3-1, Hongo, Bunkyo-ku, Tokyo, JAPAN,
(Currently belonging to NTT DATA Corporation.) so.hiral.16@gmail.com,
[2]Graduate School of Information Science and Engineering, The University of Tokyo,
7-3-1, Hongo, Bunkyo-ku, Tokyo, JAPAN, yamanishi@mist.i.u-tokyo.ac.jp

ABSTRACT

We are concerned with the issue of detecting changes of clustering structures from multivariate time series. From the viewpoint of the minimum description length (MDL) principle, we introduce an algorithm that tracks changes of clustering structures so that the sum of the code-length for data and that for clustering changes is minimum. Here we employ a Gaussian mixture model (GMM) as representation of clustering, and compute the code-length for data sequences using the normalized maximum likelihood (NML) coding. The introduced algorithm enables us to deal with clustering dynamics including merging, splitting, emergence, disappearance of clusters from a unifying view of the MDL principle. We empirically demonstrate using artificial data sets that our proposed method is able to detect cluster changes significantly more accurately than an existing statistical-test based method and AIC/BIC-based methods. We further use real customers' transaction data sets to demonstrate the validity of our algorithm in market analysis.

1. SUMMARY

1.1. Problem Setting

This paper is organized as a brief summary of our recent paper [1]. We address the issue of clustering multi-variate data sequences. Suppose that the nature of data changes over time. We are then specifically interested in tracking changes of clustering structures, which we call *clustering change detection*. We are concerned with the situation where time series data are sequentially given and the clustering must be conducted in a sequential fashion. The main purpose of this talk is to introduce, according to our recent work [1], a novel clustering change detection algorithm in the sequential setting. We employ a Gaussian mixture model (GMM) as a representation of clustering and design the algorithm on the basis of the minimum description length (MDL) principle [2]. That is, it tracks changes of clustering structures so that the sum of the code-length for data and that for clustering changes is minimum.

1.2. Previous Works

There exist a number of methods for tracking changes of clustering structures. For example, Song and Wang [3] proposed a statistical-test based algorithm for dynamic clustering. It estimates a GMM in an on-line manner and then conducts a statistical test to determine whether a new cluster is identical to an old one or not. If it is, the new cluster is merged into the older one, otherwise it is recognized as a cluster which has newly emerged. Sato [4] proposed an algorithm for merging and splitting of clusters in a GMM based on the variational Bayes method. Note that changes of clusters are not necessarily classified into merging or splitting. Siddiqui et.al.[5] proposed a method of tracking clutering changes using the EM algorithm and Kalman filters. Our work is different from Siddiqui et.al.'s one in that the former is concerned with changes of the number of clusters while the latter is concerned with parameter trajectories keeping the number of clusters fixed.

1.3. Novelty of Our Approach

The novelty of the approach in [1] may be summarized as follows:

1)*An extension of DMS into a sequential clustering setting:* Yamanishi and Maruyama [6, 7] developed a theory of dynamic model selection (DMS) for tracking changes of statistical models on the basis of the MDL principle. We extend DMS to the sequential setting to introduce a *sequential DMS algorithm* [1]. Every time data is input, it sequentially detects changes of clustering structures on the basis of the MDL principle so that the sum of the code-length for the data and that for the clustering change is minimum. This algorithm enables us to deal with the dynamics of clustering structures, including "merging", "splitting", "emergence", "disappearance", etc. within a unified framework from the viewpoint of the MDL principle.

2)*A new application of the NML code-length to sequential DMS:* In the sequential DMS algorithm,it is crucial how to choose a method for coding. The best choice is the NML coding since it has turned out to be the optimal code-length in the sense of minimax criterion [2]. However, the normalization term diverges for a multi-dimensional Gaussian distribution and it is computationally difficult to straightforwardly compute the NML code-length for a GMM exactly. Hirai and Yamanishi proposed a method for efficiently computing the NML code-length for GMMs [8], inspired by Kontkanen and Myllymäki's work [9] in which the the efficient computation of the NML code-lengths for discrete distributions was addressed. They recently modified their method using the renormalizing technique as in [10], to develop an efficient method for computing the renormalized maximum likelihood code-length (RNML) for a GMM [11]. We employ the RNML coding for GMMs in the computation process of the sequential DMS. This is the first work on the usage of the RNML coding in the scenario of sequential clustering change detection.

3)*Empirical demonstration of the superiority of the sequential DMS with the RNML code-length over the existing methods:* Using artificial data sets, we empirically demonstrate the validity of our method in comparison with Song and Wang's method [3], AIC (Akaike's information criteria)[12] / BIC (Bayesian information criteria)[13]-based tracking methods etc. We also use a real data set consisting of customers' purchase records for a number of kinds of beers. Tracking changes of clusters of customers leads to the understanding of how customers' purchase patterns change over time and how customers move from clusters to clusters. This demonstrates the validity of our method in the area of marketing.

2. ACKNOWLEDGMENTS

This work was supported by MEXT KAKENHI 23240019, Aihara Project, the FIRST program from JSPS, initiated by CSTP, NTT Corporation.

3. REFERENCES

[1] S.Hirai and K. Yamanishi, "Detecting changes of clustering structures using normalized maximum likelihood coding," *Proc. of KDD2012*, 2012.

[2] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. on Inf. Theory*, vol. 42(1), pp. 40–47, January 1996.

[3] M. Song and H. Wang, "Highly efficient incremental estimation of Gaussian mixture models for online data stream clustering," *Intelligent Computing: Theory and Application*, 2005.

[4] M. Sato, "Online model selection based on the variational bayes," *NC*, vol. 13, pp. 1649–1681, 2001.

[5] Z.F.Siddiqui G.Krempl and M.Spiliopoulou, "Online clustering of high-dimensional trajectories under concept drift," *Proc. of ECML-PKDD2011, Part II*, pp. 261–276, 2011.

[6] K. Yamanishi and Y. Maruyama, "Dynamic syslog mining for network failure monitoring," *Proc. of KDD2005*, pp. 499–508, 2005.

[7] K. Yamanishi and Y. Maruyama, "Dynamic model selection with its applications to novelty detection," *IEEE Trans. on Inf. Theory*, vol. 53, no. 6, pp. 2180–2189, June 2007.

[8] S. Hirai and K. Yamanishi, "Normalized maximum likelihood coding for exponential family with its applications to optimal clustering," *arXiv 0474364*, 2012.

[9] P. Kontkanen and P. Myllymäki, "A linear time algorithm for computing the multinomial stochastic complexity," *Information Processing Letters*, vol. 103, pp. 227–233, 2007.

[10] J. Rissanen, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, November 2000.

[11] S. Hirai and K. Yamanishi, "Efficient computation of normalized maximum likelihood coding for Gaussian mixtures with its applications to optimal clustering," *Proc. of ISIT*, pp. 1031–1035, 2011.

[12] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.

[13] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics 6 (2)*, pp. 461–464, 1978.