IDIAP RESEARCH REPORT

DICHOTOMY BETWEEN CLUSTERING PERFORMANCE AND MINIMUM DISTORTION IN PIECEWISE-DEPENDENT-DATA (PDD) CLUSTERING

Itshak Lapidot Hugo Guterman

IDIAP-RR 02-48

Dalle Molle Institute
for Perceptual Artificial
Intelligence • P.O.Box 592 •
Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

Dr. Hugo Guterman - head of Neural Networks and Fuzzy Logic Lab. at the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Israel.

IDIAP Research Report 02-48

DICHOTOMY BETWEEN CLUSTERING PERFORMANCE AND MINIMUM DISTORTION IN PIECEWISE-DEPENDENT-DATA (PDD) CLUSTERING

Itshak Lapidot Hugo Guterman

OCTOBER 2002

Abstract. In many signal such speech, bio-signals, protein chains, etc. there is a dependency between consecutive vectors. As the dependency is limited in duration such data can be called as Piecewise-Dependent-Data (PDD). In clustering it is frequently needed to minimize a given distance function. In this paper we will show that in PDD clustering there is a contradiction between the desire for high resolution (short segments and low distance) and high accuracy (long segments and high distortion), i.e. meaningful clustering.

Index Terms: clustering, minimal distance, self-organizing maps, piecewise-dependent-data.

1. Introduction

PDD clustering has many applications in time-signals including speaker recognition [1]-[7]; machine monitoring [8]; clustering of EEG signals [9]; and music clustering [10]. Similar methods have also been applied in other areas, such as protein modeling [11].

PDD clustering must be used when there is a successive dependence between a group of data vectors and there no trained models of the data classes. In distance-measure-based clustering the goal is usually to minimize the overall distance. When a PDD is used longer segment supply more information about the clusters and better clustering can be performed [1], [2], [5]. The desires for minimum distance and maximum segment length are contradict between themselves. To achieve minimum distance smaller segments must be used (one feature vector per segment will lead to best minimization) but the clusters may be meaningless, large segments will lead to better clustering results but high distance. In other words the PDD clustering criterion should be as follows: minimize the overall distance under the restriction of the largest segments that can be used.

The rest of the paper is as follows: problem formulation and the proof is given in section II. In section III PDD clustering algorithm we used is described. Experiment and results are presented in section IV. Section V summarized the paper.

2. Problem Formulation and Proof

If the data $\mathbf{V}=\{v_n\}_{n=1,\ldots,N}$ consists of M segments, as described in equation 1, and the data have to be clustered into R clusters ($N \gg R$) such that vectors in the same segment must not be separated. Then the best clustering will be obtained if $M=R$, i.e., all the data consists of R segments, each segment representing a different cluster. The worst case is when $M=N$, i.e., each vector is a segment and there is almost no information in each segment about cluster statistics. We will show that in this case a better local minimum may be achieved.

$$\begin{cases} \mathbf{V}=\left\{\underbrace{v_1^1,\ldots,v_{n_1}^1}_{\mathbf{V_1}},\ldots,\underbrace{v_1^m,\ldots,v_{n_m}^m}_{\mathbf{V_m}},\ldots,\underbrace{v_1^M,\ldots,v_{n_M}^M}_{\mathbf{V_M}}\right\}=\{\mathbf{V_m}\}_{m=1,\ldots,M} \\ \sum_{m=1}^{M} n_m = N \end{cases} \tag{1}$$

Two cases will be presented to prove that smaller segments may reach a lower distance minimum value. First, in subsection A, a special case will be proved for when two partitions are the same, with the exception of $1 \le m \le M$ segments that split into two or more sub-segments in one of the partitions. In subsection B, a general case will be proved from a probabilistic point of view. It will be shown that if there are two different partitions ($\{\mathbf{V_m^1}\}$ and $\{\mathbf{V_m^2}\}$) with M_1 and M_2 segments respectively ($M_1 < M_2$), there is a greater probability to reach a lower distortion with $\{\mathbf{V_m^2}\}$ than with $\{\mathbf{V_m^1}\}$.

2.1 Special Case

With no limitation of generality assume that:

$$\begin{cases} \mathbf{V}_{\mathbf{m}}^{\mathbf{1}} = \mathbf{V}_{\mathbf{m}}^{\mathbf{2}} & ; \quad 1 \le m \le M_1 - 1 \\ \left[\mathbf{V}_{\mathbf{M_1}}^{\mathbf{2}} \quad \mathbf{V}_{\mathbf{M_2}}^{\mathbf{2}} \right] = \mathbf{V}_{\mathbf{M_1}}^{\mathbf{1}} & ; \quad M_2 = M_1 + 1 \end{cases} \tag{2}$$

and that the labeling of $\left\{ \mathbf{V}_{\mathbf{m}}^{\mathbf{1}} \right\}$ is

$$\begin{cases} \left\{ r_m^1 \right\}_{m=1,\dots,M_1} \\ 1 \le r_m^1 \le R \end{cases} . \tag{3}$$

If the labeling of $\left\{ \mathbf{V}_{\mathbf{m}}^{\mathbf{2}} \right\}$ is

$$\begin{cases} r_m^2 = r_m^1 & ; \quad 1 \le m \le M_1 \\ r_{M_2}^2 = r_{M_1}^1 \end{cases} \tag{4}$$

then the distortions of the systems with $\left\{ \mathbf{V}_{\mathbf{m}}^{\mathbf{1}} \right\}$ and $\left\{ \mathbf{V}_{\mathbf{m}}^{\mathbf{2}} \right\}$ are the same, as $\left[\mathbf{V}_{\mathbf{M_1}}^{\mathbf{2}} \quad \mathbf{V}_{\mathbf{M_2}}^{\mathbf{2}} \right] = \mathbf{V}_{\mathbf{M_1}}^{\mathbf{1}}$.

Assuming that the labeling of $\mathbf{V}_{\mathbf{M_1}}^{\mathbf{2}}$ and $\mathbf{V}_{\mathbf{M_2}}^{\mathbf{2}}$ are $r_{M_1}^2$ and $r_{M_2}^2$, respectively, and $r_{M_1}^2 \ne r_{M_2}^2$, then the partition of $\left\{ \mathbf{V}_{\mathbf{m}}^{\mathbf{1}} \right\}$ is not valid for $\left\{ \mathbf{V}_{\mathbf{m}}^{\mathbf{2}} \right\}$. This means that if one of the partitions with $r_{M_1}^2 \ne r_{M_2}^2$ gives a lower distance than any of the partitions with $r_{M_1}^2 = r_{M_2}^2$, a clustering using $\left\{ \mathbf{V}_{\mathbf{m}}^{\mathbf{2}} \right\}$ may give a lower distance than a clustering using $\left\{ \mathbf{V}_{\mathbf{m}}^{\mathbf{1}} \right\}$.

2.2 General Case

In the general case there are no assumptions about the data partition. The only assumption, with no limitation of generality, about $\left\{ \mathbf{V}_{\mathbf{m}}^{\mathbf{1}} \right\}$ and $\left\{ \mathbf{V}_{\mathbf{m}}^{\mathbf{2}} \right\}$ is that $M_1 < M_2$.

If the length of all the segments is equal to one, then the number of segments is $M = N$ and the number of different partitions between R clusters is $R^M = R^N$. The number of different partitions of $\left\{ \mathbf{V}_{\mathbf{m}}^{\mathbf{1}} \right\}$ and $\left\{ \mathbf{V}_{\mathbf{m}}^{\mathbf{2}} \right\}$ are R^{M_1} and R^{M_2} respectively, and the inequality $R^{M_1} < R^{M_2}$ holds. Consequently, the probabilities of $\left\{ \mathbf{V}_{\mathbf{m}}^{\mathbf{1}} \right\}$ and $\left\{ \mathbf{V}_{\mathbf{m}}^{\mathbf{2}} \right\}$ to achieve global minimum are $P_1 = {R^{M_1}}/{R^M}$ and $P_2 = {R^{M_2}}/{R^M}$ respectively. As R^{M_2} is greater than R^{M_1}, the probability P_2 is greater than P_1, i.e., the probability to reach a lower distortion increases with the number of segments.

3. PDD Clustering Algorithm

The following describes algorithm we already used in [1] and [2]. In general, given a PDD the goal is to cluster the data into R clusters. The PDD consists of N vectors, $\mathbf{V} = \left\{ v_n \right\}_{n=1,\dots,N}$. These vectors are partitioned into M segments, $\mathbf{V} = \left\{ \mathbf{v}_{\mathbf{m}} \right\}_{m=1,\dots,M}$ (equation 1). The segments have to be clustered into R clusters, such that two vectors that belong to the same segment must be clustered to the same cluster. A CB is created, for each model, using distance-measure-based algorithm.

The initiation of the process is performed by randomly assigning equal number of segments to all CB_r s ($\mathbf{V}^{\mathbf{r},\mathbf{0}}$-segments that partitioned to CB_r at the beginning). Each model is trained using the data assigned to it during the partitioning. After the training the regrouping process is applied and a new segment attribution is given according to the minimal distance. The regrouping process produces a new partition and the models are retrained again. Hence, an iteration of the clustering process is defined as follows:

1. Retrain the models with the new partition achieved by the previous iteration.
2. Regroup the data by finding minimal distance between each segment and the retrained models.
3. Test for termination: if the termination criterion is met, exit; if not return to 1.

For models training can use any distance-based-algorithm algorithm that converges at least to a local minimum, such as the LBG [12], Self-Organizing Map (SOM) [13], fuzzy C-means [14] etc.

At the end of this iterative procedure R models for the R clusters are provided. The data is segmented and labeled. A proof of the algorithm convergence can be found in [1]. The present system employs SOM [14] for CB production.

4. Experiments and Results

This experiment shows that selection of short segments can produce incorrect clusters even when the overall distance is low.

The following time-series, composed of two models, is assumed. The output of the time-series is taken from one of the two models. The pdfs of the model, $f_x(\alpha)$ and $f_y(\beta)$, are:

$$\begin{cases} f_x(\alpha) = \dfrac{1}{(2\pi)^{\frac{1}{2}}} \exp\left\{ \dfrac{-(\alpha)^2}{2} \right\} \\ f_y(\beta) = 0.2 \cdot \dfrac{1}{(2\pi)^{\frac{1}{2}}} \exp\left\{ \dfrac{-(\beta+2)^2}{2} \right\} + 0.8 \cdot \dfrac{1}{0.5 \cdot (2\pi)^{\frac{1}{2}}} \exp\left\{ \dfrac{-(\beta-10)^2}{2 \cdot (0.5)^2} \right\} \end{cases} \tag{5}$$

and the pdfs of both models are shown in Fig. 1.

A time-series of 2000 samples was generated. The models switched every fifty samples. Two tests were produced for estimation of the models' clusters. The segment lengths for the first and second test were fifty samples and one sample, respectively. A SOM of size 1×3 was used for each model. The results are shown in Fig. 2. Fig. 2a is the histogram of the first test and it can be seen that it fits the generated models (Fig. 1). The results were always consistent and similar. Fig. 2b-d shows the histograms for the second test. It can be seen that the SOMs converged, each time, to a different local minimum that dose not fits the original models (Fig. 1). However, the overall Euclidean distance is always lower than it was in the first test.

5. Conclusions

Since, in clustering problem minimal distance is not a goal but a way to achieve the goal, it is better to use the largest segments that are possible. Large segments can ensure a sufficient statistic for good clustering. The distance-based algorithm should be applied under the restriction of sufficient segment length. From the experiment we saw that short segments leads to low distance result, as can be expect from the proof in section II, and meaningless clusters, while large segments leads to correct cluster despite the fact that the overall distance was very high.

Knowledge or lack of knowledge about the boundaries of each data segment influences the problem's complexity. In several applications the segmentation is already given and only the labeling is missing [10], [11] or the segmentation can be found in advance [8].

Consequently, in these cases all the available data for each segment should be employed. In other cases segment boundaries are unknown but might be estimated by combining a minimal duration constrain to ensure a sufficient statistics of each segment and a Viterbi search to find the correct segmentation [3]. This procedure is done every time instead of step 2 in the clustering algorithm.

In a previous work [1] several tests were performed to find the influence of segments length on speaker clustering performances. It was found that short segments (50 input vectors per segment) need longer training and achieved high error rate. Long segments (200 input vectors per segments) contain too many segments that have data of several speakers (split segments). Such segments can be noisy from the clustering point of view and lead to high error rate as well. In such case the optimal segment length should be estimated or found empirically for each application.

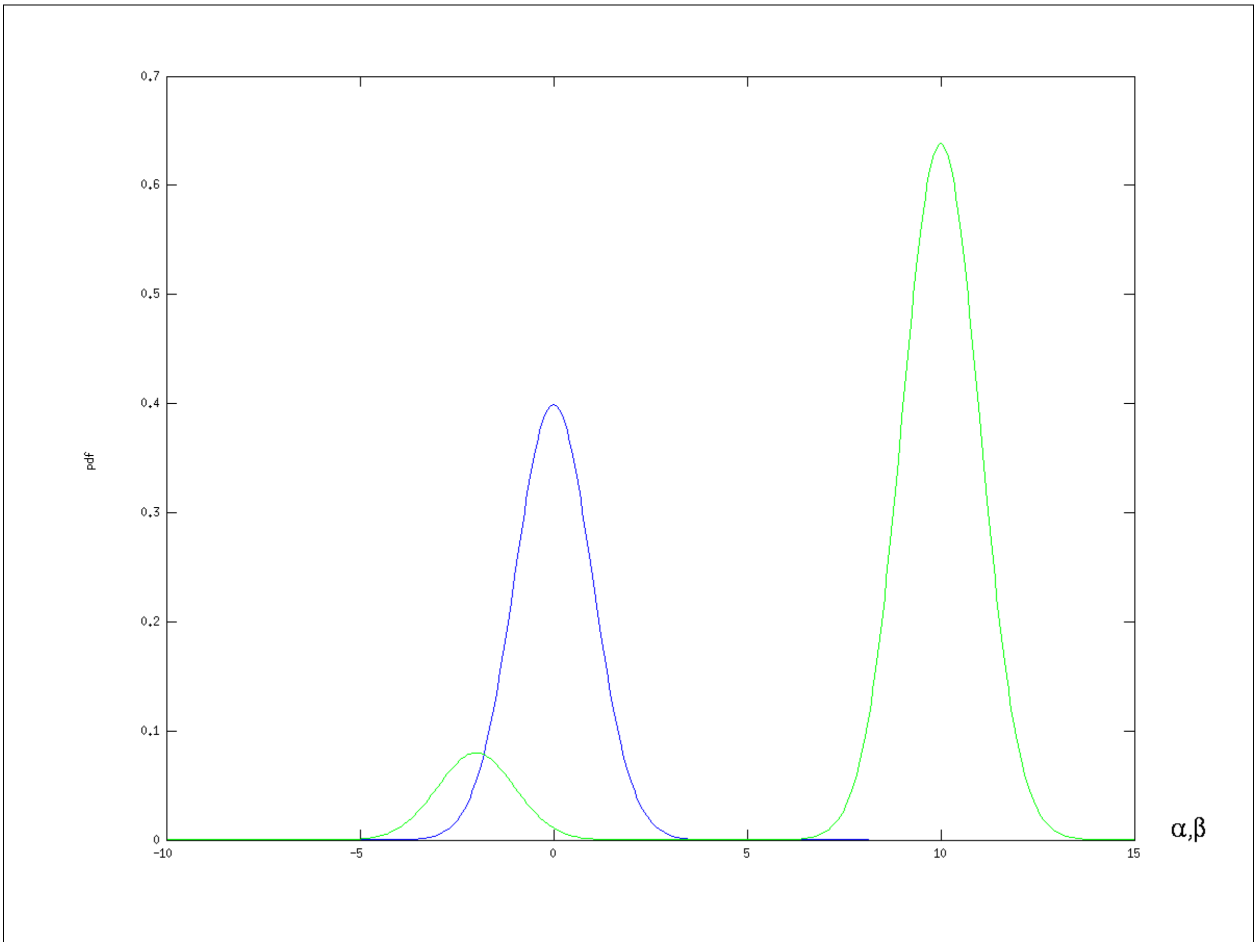

Fig 1: The two models pdfs: green line – $f_x(\alpha)$, blue line – $f_y(\beta)$.

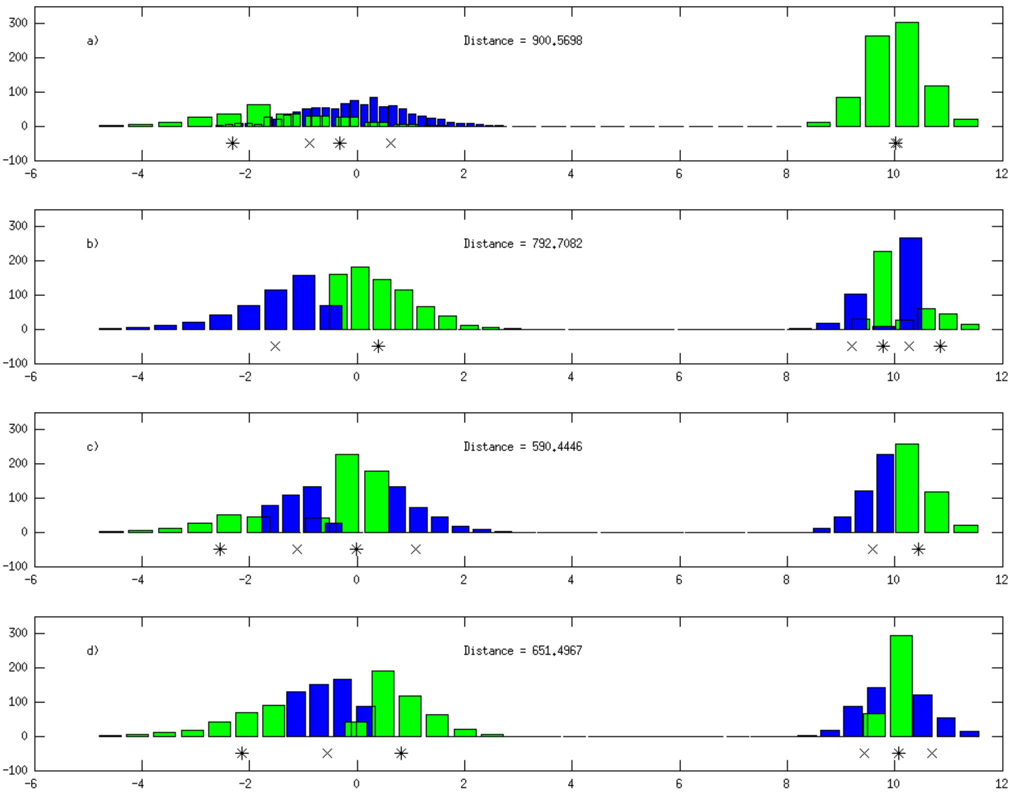

Fig. 2: Histograms of the clustering results: blue – 1st cluster, green – 2nd cluster. The places of the SOMs CW s: crosses – 1st SOM, stars – 2nd SOM.

Reference

[1] I. Lapidot (Voitovetsky), H. Guterman, and A. Cohen, "Unsupervised Speaker Recognition Based on Competition Between Self-Organizing-Maps," *IEEE Trans. on Neural Networks*, vol. 13, no. 4, pp. 877-887, July 2002.

[2] I. Lapidot (Voitovetsky) and H. Guterman "Resolution Limitation in Speakers Clustering and Segmentation Problems," *Proc. 2001 A Speaker Odyssey*, 2001.

[3] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, "Unknown-multiple speaker clustering using HMM," *Proc. International Conference on Spoken Language Processing*, pp. 573-576, September 16-20, 2002, Denver, Colorado, USA.

[4] A. Cohen and V. Lapidus, "Unsupervised, text independent, speaker classification," *Proc. of the Int. Conf. on Signal Processing Application and Technology*, pp. 1745-1749, 1996.

[5] M. Sugiyama, J. Murakami, and H. Watanabe, "Speech segmentation and clustering based on speaker features," *Proc. International Conference on Acoustic Speech and Signal Processing*, vol. 2, pp. 395-398, 1993.

[6] H. Gish, M.-H. Siu, and R. Rohlicek, "Segregation of speaker for speech recognition and speaker identification," *Proc. International Conference on Acoustic Speech and Signal Processing*, vol. 1, pp. 873-876, 1991.

[7] J. O. Olsen, “Separation of speakers in audio data,” *Proc. of 4th European Conference on Speech Communication and Technology,* vol. 1, pp. 355-358, 1995.

[8] L. M. D. Owsley, L. E. Atlas, and G. D. Bernard, “Self-organizing feature maps and hidden Markov models for machine-toll monitoring,” *IEEE Trans. on Signal Processing*, vol. 45, no. 11, pp. 2787-2798, November 1997.

[9] J. Kohlmorgen, K.-R. Muller, and K. Pawelzik, “Segmentation and identification of drifting dynamical systems,” *Proc. Neural Networks for Signal Processing VII IEEE Workshop*, pp. 326-335, 1997.

[10] O .A. S. Carpinteiro, “A hierarchical self-organising map model for sequence recognition,” *Pattern Analysis and Applications*, vol. 3, no. 3, pp. 289-287, 2000.

[11] A. Krogh, M. Brown, I. Saira Mian, K. Sjolander, and D. Haussler, “Hidden Markov models in computational biology applications to protein modeling,” *J. Mol. Biol.*, vol. 235, no. 5, pp. 1501-1531, 1994.

[12] Y. Linde, A. Buzo, and R. M. Gray, “An algorithm for vector quantizer design,” *IEEE Trans. Commun.*, vol. COM-28, no. 1, pp. 84-95, January 1980.

[13] T. Kohonen, “The self-organizing map,” *Proc. IEEE*, vol. 78, no. 9, pp. 1464-1480, September 1990.

[14] W. Pedrycz, “Fuzzy sets in pattern recognition: methodology and methods,” *Pattern Recognition*, vol. 23, no. 1/2, pp. 121-146, 1990.