Automated Sequence Clustering of Audio Signals using Conditional Random Fields

Dierck Matern, Alexandru Condurache and Alfred Mertins

University of Luebeck, Ratzeburger Allee 160, 23562 Luebeck, Germany
{*matern, condurache, mertins*}*@isip.uni-luebeck.de*

Abstract

In this paper, we propose a new conditional random field (CRF) based algorithm for automated sequence clustering of audio signals, where the term *sequence clustering* is used in the same way as in biomedical signal clustering. Usual CRF based methods are trained in an observed manner, that is, for the training data, we need a corresponding state sequence. For automated sequence clustering, we have no such known state sequence in the beginning. We therefore adapt the training of the CRFs to the case of an unknown state sequence and apply the trained model for classification of new instances of the audio signals. This analysis of the acoustical signals can be used, for example, for scene analysis or novelty detection, where one uses abstract states regularly and therefore manual prelabeling is not reasonable. In the experiments, we successfully have applied the trained model for sequence clustering to audio signals and were able to detect the significant clusters.

Introduction

Sequence clustering [8] is an interesting tool for preprocessing of different signal analysis applications. We may wish to divide the whole signal into parts by grouping the features according to common properties, and, for example, apply individual classification on the clusters. Those classifications can be to detect events [3, 4] in the whole sequence, in parts of the sequence (for example, disturbances in a speech signal which can be ignored in silent parts), comparison of different audio files or context information in natural language analysis. The grouping of features is called "sequence clustering" in this paper.

The goal of the paper is to demonstrate that conditional random fields (CRFs) can be applied for blind sequence clustering in audio signals. CRFs are Markov models which have been applied for text analysis [2] and other signal processing problems [4]. We propose an unsupervised training algorithm, which means that we can apply the clustering in a blind and fully automated manner. We want the clustering to fulfill two criteria: first, the state sequence has to provide high information (that is, the entropy of the state sequence is high) and second, we want each cluster to be identified as good as possible (that is, the likelihood of one state at one time step is high while it is low for all other states). Therefore, we maximize the entropy of the generated state sequence while minimizing the entropy at each time step in the training of the CRF.

In the rest of the paper, we first discuss the training algorithm. After that, we show in some experiments the clustering applied to an audio signal as an example of the algorithm. We conclude the paper with a discussion.

Sequence Clustering using Conditional Random Fields

CRFs are graphical models [5, 7] which are often used to label sequences [2]. We can calculate the posterior probabilities of a state sequence $\mathbf{S} = s(1), s(2), \ldots, s(N)$ given the feature sequence $\mathbf{X} = \mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)$, that is $p(\mathbf{S}|\mathbf{X};\lambda)$ where λ is the parameter vector of the CRF [2]. We can also estimate a state sequence by identifying the one with the highest likelihood, $\hat{\mathbf{S}} = \arg\max_{\mathbf{S}} p(\mathbf{S}|\mathbf{X};\lambda)$. Each state ζ_k with $k = 1, 2, \ldots, K$ represents a cluster, that is, for each n we have $s(n) \in \{\zeta_1, \zeta_2, \ldots, \zeta_K\}$.

The (linear chain) CRF is defined as [2]

$$p(\mathbf{S}|\mathbf{X};\lambda) = \frac{1}{Z_\lambda(\mathbf{X})} \exp\left(\sum_{n=1}^{N}\sum_{k=1}^{K} \lambda_k^\top \varphi_k\left(s(n), \mathbf{x}(n)\right)\right) \cdot \exp\left(\sum_{n=1}^{N}\sum_{j,k=1}^{K} \lambda_{jk}\varphi_{jk}(s(n-1), s(n))\right) \tag{1}$$

where $\{\lambda_i, \lambda_{jk}\} \in \lambda$ with $i, j, k = 1, 2, \ldots, K$ is the set of the parameters for the CRF, $Z_\lambda(\mathbf{X})$ is a normalization constant such that $p(\mathbf{S}|\mathbf{X};\lambda)$ is a probability, and

$$\varphi_k(s(n), \mathbf{x}(n)) = [\![s(n) = \zeta_k]\!] \cdot \mathbf{x}(n),$$
$$\varphi_{jk}(s(n-1), s(n)) = [\![s(n-1) = \zeta_j]\!] \cdot [\![s(n) = \zeta_k]\!],$$

where $[\![P]\!]$ is 1 if P is true and 0 else.

The usual training algorithms [1, 2] consider a known state sequence. Because we want a blind clustering (that is, we do not have a state sequence for the training), we estimate a state sequence in the iterations of the training.

The state sequence $\tilde{\mathbf{S}}$ for the training is calculated using

$$\tilde{p}(s(n) = \zeta_k|\mathbf{X}, s(n-1), s(n+1);\lambda) = \frac{p(s(n) = \zeta_k|\mathbf{X}, s(n-1), s(n+1);\lambda)}{\sum_{n=1}^{N} p(s(n) = \zeta_k|\mathbf{X}, s(n-1), s(n+1);\lambda)},$$

the normalized posterior probability, in order to keep the entropy of the state sequence as high as possible. Note that $\tilde{p}$ is not a probability, but we use the same notation for consistence. $\tilde{\mathbf{S}}$ is calculated by $\tilde{\mathbf{S}} = \arg\max_{\mathbf{S}} \tilde{p}(\mathbf{S}|\mathbf{X};\lambda)$.

Further, we use a special regularization for the optimization in the training phase that is different from other CRF algorithms [1]. Because we perform an Expectation-Maximization-like training algorithm [6], we need to prevent an overfitting of

the CRF in a more comprehensive manner than other CRF algorithms. For this, we use the constraints

$$\sum_{i=1}^{M}(\lambda_j(i))^2 = 1, \qquad j = 1, 2, \ldots, K, \tag{2}$$

$$\sum_{k=1}^{K}\lambda_{jk} = 0, \qquad j = 1, 2, \ldots, K, \tag{3}$$

$$\sum_{k=1}^{K}(\lambda_{jk})^2 = 1, \qquad j = 1, 2, \ldots, K \tag{4}$$

in the training, instead of the usual application of a penalty term [1, 2]. After the optimization of the CRF, we re-calculate the state sequence and optimize again. Iteratively, the training converges to a CRF that fulfills our criteria we stated in the introduction.

For the experiments, we estimate the sequence for which each state has the highest likelihood $\hat{\mathbf{S}}$. Because we have used a state sequence with high information in the training, the entropy of $\hat{\mathbf{S}}$ is also high.

Experiments

In the experiments, we apply the clustering to the spectrum of an audio signal. These are initial experiments which are supposed to indicate whether the method is useful for real-world problems.

We calculate a spectrogram from the audio signal (an interpretation of Bach - Air, sampled at 44100Hz) with a Hanning window with a size of 256 samples and an overlap of 128 samples. We normalize each band with respect to mean and variance. We train a CRF on the first 20% of the data and evaluate on the whole sequence. We test the sequence with the highest likelihood, $\hat{\mathbf{S}} = \arg\max_{\mathbf{S}} p(\mathbf{S}|\mathbf{X};\lambda)$. As an accuracy measure, we use

$$v(\hat{\mathbf{S}},\lambda) = H(\hat{\mathbf{S}}) \cdot \left(\prod_{n=1}^{N} K \cdot p(\hat{s}(n)|\mathbf{x}(n);\lambda)\right)^{\frac{1}{N}}, \tag{5}$$

which is a combination of the similarity measure used in [8], adapted to CRFs and normalized such that it is invariant to the length of the sequence, and the entropy of the state sequence $H(\hat{\mathbf{S}}) = -\sum_{k=1}^{K} h(\hat{\mathbf{S}},\zeta_k)\cdot\log(h(\hat{\mathbf{S}},\zeta_k))$ with $h(\hat{\mathbf{S}},\zeta_k) = \frac{1}{N}\sum_{n=1}^{N}[\![\hat{s}(n)=\zeta_k]\!]$. Therefore, $v(\hat{\mathbf{S}},\lambda)$ is a measure for the claims we have for a good clustering. We measure the accuracy for the clear signal, the clear signal with a short disturbance (a siren with less loudness then the original signal, not included in the training), and, for comparison purposes, a pure random signal. We have repeated the experiments for each parameter five to ten times, with random initializations. The results were the same each time.

As can be seen in Table 1, the difference of the clustering produced by the disturbance is visible, but not significant; that is, the method is robust to disturbances. With a high number of states K, we can identify the disturbance. The noise can be clearly distinguished. Because of the high values compared to the pure noise, we rate the experiments as a success, that is, we could clearly identify the significant parts of the audio signal.

Table 1: Results of given experiments with different numbers of state K. Denoted is the accuracy measure in Equation (5) for the clean signal, a disturbed signal and pure noise.

	$K=2$	$K=3$	$K=4$	$K=6$
Clean signal	1.8443	2.6956	3.4150	4.9168
Disturbed signal	1.8583	2.6843	3.3772	4.9089
Noise	1.5508	1.8968	2.1725	2.5740

Discussion

We have proposed a blind sequence clustering algorithm based on CRFs. The results of the experiments, which are an initial analysis of the method, fulfilled our expectations.

Especially, this method can be a first step for the event detection in audio streams. Because those events may depend on the current situation of the audio signal, the clustering can be important. For example, the event can be detected if we consider local properties only (for example, disturbances) or we may need global information (for example, a wrong order in a repetitive signal). Hence, the clustering of audio signals can be a great improvement for signal analysis tasks.

References

[1] Rahul Gupta. Conditional random fields. Technical report, IIT Bombay, 2006.

[2] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001.

[3] D. Matern, A. P. Condurache, and A. Mertins. Linear Prediction based Mixture Models for Event Detection in Video Sequences. In *Proc. Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA) 2011*, pages 25 – 32, Gran Canaria, Spain, 2011.

[4] D. Matern, A. P. Condurache, and A. Mertins. Event detection using log-linear models for coronary contrast agent injections. In *Proceedings of the First International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, volume 2, pages 172–179, Vilamoura - Algarve, Portugal, 2012.

[5] A. McCallum, D. Freitag, and F. C. N. Pereira. Maximum entropy Markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 591–598, San Francisco, CA, USA, 2000.

[6] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2. edition, 2008.

[7] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. volume 77, pages 257–286. Feb 1989.

[8] J. Yang and W. Wang. CLUSEQ: Efficient and effective sequence clustering. In *Proceedings of the 19th International Conference on Data Engineering*, pages 101–112. IEEE Press, 2003.