International Journal of Advanced Smart Convergence Vol.7 No.2 33-37 (2018)
http://dx.doi.org/10.7236/IJASC.2018.7.2.33

IJASC 18-2-4

Determining the Optimal Number of Signal Clusters Using Iterative HMM Classification

Duker Ernest Junior*, Yoon Joong Kim**

Department of Computer Engineering, Hanbat National University, Korea,**
dukerernestjunior@gmail.com,yjkim@hanbat.ac.kr***

Abstract

In this study, we propose an iterative clustering algorithm that automatically clusters a set of voice signal data without a label into an optimal number of clusters and generates hmm model for each cluster. In the clustering process, the likelihood calculations of the clusters are performed using iterative hmm learning and testing while varying the number of clusters for given data, and the maximum likelihood estimation method is used to determine the optimal number of clusters.

We tested the effectiveness of this clustering algorithm on a small-vocabulary digit clustering task by mapping the unsupervised decoded output of the optimal cluster to the ground-truth transcription, we found out that they were highly correlated.

Keywords: *HMM, Iterative Clustering, Optimal, Unsupervised, Signal.*

1. Introduction

The area of speech processing, which aims to develop unsupervised ways to learn and cluster speech data in environments where label manuscripts are not available, has received tremendous attention. Different clustering approaches have been proposed in the literature, most of which consist of two processes; first is the segmentation of the signal into homogeneous segments and the other is the grouping of the segments. The segmentation is either assumed to be known [1-2] or performed automatically before clustering [3,6-8]. These approaches have limitations: in the former case, the correct segmentation is hardly known before the practical applications, and in the latter case, the errors made in the segmentation step are not only difficult to correct later, but can propagate into the subsequent clustering step and thereby affecting performance.

The framework of unsupervised training based on Hidden Markov Model (HMM) was a recently proposed approach, that has shown to be useful for building speech recognizers where there are no transcriptions available [3-4]. A recent approach by [3] used unsupervised HMM training approach as an optimization over both the parameters of training and transcription sequence to produce self-organizing units. The study [4] used the similar approach of [3] with application to topic classification and keyword discovery. Both

Manuscript Received: Apr. 20, 2018 / Revised: Apr. 30, 2018 / Accepted: May. 9, 2018
Corresponding Author: yjkim@hanbat.ac.kr
Tel: +82-42-821-1143, Fax: +82-42-821-1595
Department of Computer Engineering, Hanbat National University, Korea

researchers recorded remarkable results, as this seems to prove the effectiveness of Unsupervised HMM-training on unlabeled speech data. In both approach, segmentations are done prior to training and clustering.

In this study, we propose an iterative clustering algorithm that automatically clusters a set of voice signal data without labeling into an optimal number of clusters and generates hmm model for each cluster. In the clustering process, the likelihood calculations of the clusters are performed using iterative hmm learning and testing while varying the number of clusters for given data, and the maximum likelihood estimation method is used to determine the optimal number of clusters.

2. Iterative clustering algorithm

The iterative clustering algorithm is configured as shown in Figure 1. In the initialization process, the maximum number of clusters K, the training data set X, and the number of clusters being created k are initialized.

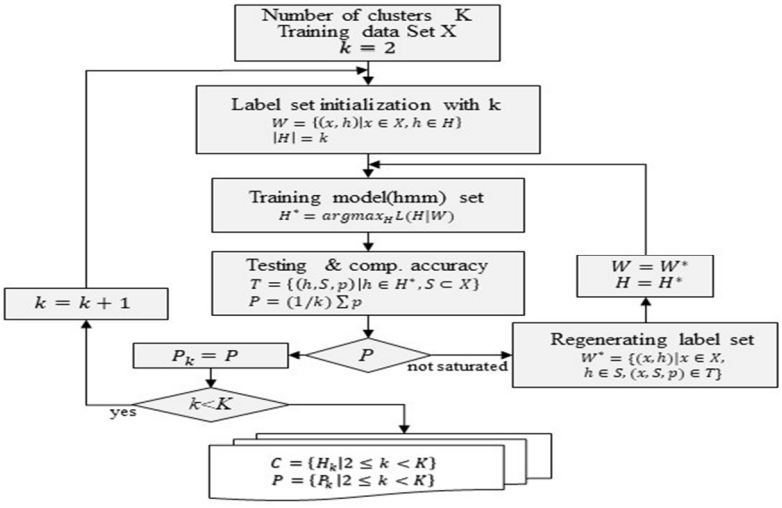

Figure 1. Block diagram of data clustering process

In the label set initialization process, a label set W composed of k (x, h) is generated. That is, h is randomly generated number between 1 and k and is assigned to training data x. This number refers to the element index of the model set H consisting of k empty models.

$$W = \{(x,h)|x \in X, h \in H\},\ |H| = k \quad (1)$$

In the training process, all hmm parameters in the hmm set H are re-estimated according to the description in the label set W and new hmm set H^* is generated using the Baum-Welch algorithm. The implementation details are presented in detail in the experimental section.

$$H^* = \text{argmax}_{\text{H}}\, P(H|W) \quad (2)$$

In the testing process, we use the Viterbi algorithm to perform recognition experiments on all data set X with the generated hmm set H^* and the label set W. For each hmm model h, the recognized data set S, and the recognition rates are obtained, and the T is a result of the recognition test as shown in equation 3. The recognition rate ps in one model set are averaged to obtain the recognition rate P of the model set H^*.

$$T = v(X|H^*, W) = \{(h, S, p)|h \in H^*, S \subset X\} \quad (3)$$

$$P = \frac{1}{k}\sum p \quad (4)$$

In the label regeneration process, if the recognition rate P is improved, a new label set W^* is generated according to the recognition result T. That is, based on the data (h, S, p) of T, the data and hmm index (x, h) are generated. H^* and W^* are stored in H and W repectively, and the process is performed again from the training process.

$$W^* = \{(x, h)|x \in X,\ h \in S, (x, S, p) \in T\} \quad (5)$$

If the recognition rate P is no longer improved, the P is stored in P_k as a likelihood value of the hmm set of size k. If the number of working clusters k is checked and k is less than the maximum number K of clusters, then k is incremented and the process is performed again from the training process. If k is equal to K, the process is terminated and a collection of hmm sets and a set of recognition rates for each hmm set are output.

$$C = \{H_k|2 \le k < K\} \quad (6)$$

$$P = \{P_k|2 \le k < K\} \quad (7)$$

3. Computation of optimal number of clusters

From the given data set X, a collection C of hmm sets with a set size between 4 and K is obtained and the corresponding set P of recognition rates is obtained.. Since the set P is defined as a set of likelihoods of the collection C, the hmm set having the greatest probability among P can be simply chosen as an optimal number of clusters. However, if many trials of data are accumulated, they will have the characteristics of a normal distribution, which may be chosen as the maximum likelihood estimation method.

In the former case, the index k^* of the maximum element is chosen in the recognition rate set P, and the optimal hmm set H_{k^*} is chosen using this value k^*.

$$k^* = \text{argmax}_\text{k}\, P_k \quad (8)$$

The k^* is called the optimal number of clusters and the size of the hmm set to generate the maximum recognition rate for the given data set X.

In the latter case, the recognition rate set P is assumed to be normal distribution expressed as a random variable of the size of the hmm set. A probability density function of the normal distribution is defined by a random variable k and parametrized in terms of the mean and the variance, denoted by μ and σ^2 respectively.

$$f(k; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(k-k)^2}{2\sigma^2}} \quad (9)$$

This equation can be defined as the likelihood of the size k of the hmm set. The optimal number of clusters can be determined by the maximum likelihood estimation method derived from this equation.

4. Experiment

To evaluate the proposed approach, we use 35 segment data made from 8 kHz recordings of English

pronunciations of the digits 0-9. These recordings are of varying durations, with a mean of approximately 0.5 seconds. HTK toolkit was used in this experiment.

The segment is converted to a series of MFCC(Mel Frequency Cepstral Coefficient) vectors according to the description in a configuration file which specifies the various configuration parameters such as the format of the speech file, length of time frame(25ms), frame periodicity (10s), number of MFCC coefficients(12) etc. The MFCC feature vectors are used not only in training but also in testing process.

The cluster number K is initialized to four and used in the formation of the language model (Task Grammar) and the pronunciation dictionary. The task grammar is defined using extended Backus-Naur form (EBNF) in a text file. The task grammar is compiled with HTK tool HParse to generate the task network.

The label data set is generated as a script file according to the number of clusters K and used in re-estimating hmm parameters (transition probability, mean and variance vectors for each observation) using HTK tool HRest. The recognition test is performed using HVite command of the HTK toolkit that produces recognition result used for regenerating the label set. These processing steps are iteratively performed until convergence

The table 1. shows a comparison of the statistical averages of the likelihood scores after training and testing for hmm sets of various sizes K. In this experiment, the optimal number K of models in cluster(hmm set) is simply determined as a set of highest probabilities using the method of Equation 8. We can deduce that the 5th hmm set is optimal for the given set of data as it gives the highest of the average statistical score as shown in table 1. This can also be seen in Figure 2 showing a graph of K against average score. We should indicate that the number of data sample was kept constant throughout the experiment.

Table 1. Average scores of the various K models

Cluster with size K	Average score
4 model cluster	900.2052
5 model cluster	905.1154
6 model cluster	893.6547
7 model cluster	892.6143
8 model cluster	884.7652
9 model cluster	880.8612

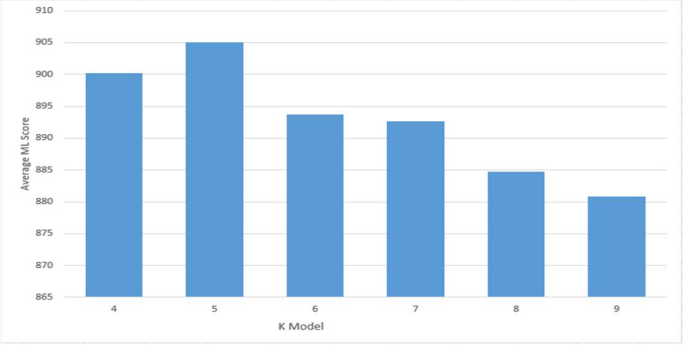

Figure 2. Score comparison of the various K models

5. Conclusion

In this study, we propose an iterative clustering algorithm that automatically clusters a set of voice signal data without labeling into an optimal number of clusters and generates hmm model for each cluster. In the

clustering process, the likelihood calculations of the clusters are performed using iterative hmm learning and testing while varying the number of clusters for given data, and the maximum likelihood estimation method is used to determine the optimal number of clusters.

The system was tested on a group of small vocabulary digit clustering. Experiments indicated the output of the selected cluster from the proposed method highly correlates with the ground-truth.

In the future, we will conduct experiments on more data and conduct research to determine the optimal number using the maximum likelihood estimation method.

References

[1] Johnson, S.E. and Woodland, P.C., "Speaker clustering using direct maximisation of the MLLR-adapted likelihood," In *Fifth International Conference on Spoken Language Processing*. 1998.

[2] Solomonoff, A., Mielke, A., Schmidt, M. and Gish, H., "Clustering speakers by their voices," *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on* (Vol. 2, pp. 757-760). IEEE, . May 1998.

[3] Siu, M.H., Gish, H., Chan, A., Belfield, W. and Lowe, S., "Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery," Computer Speech & Language, 28(1), pp.210-223., 2014.

[4] Siu, M.H., Gish, H., Chan, A. and Belfield, W., "Improved topic classification and keyword discovery using an HMM-based speech recognizer trained without supervision," Eleventh Annual Conference of the International Speech Communication Association., 2010.

[5] Le, V.B. and Besacier, L., "Automatic speech recognition for under-resourced languages: application to Vietnamese language." IEEE Transactions on Audio, Speech, and Language Processing, 17(8), pp.1471-1482, 2009..

[6] Kamper, H., Livescu, K. and Goldwater, S., "An embedded segmental k-means model for unsupervised segmentation and clustering of speech," arXiv preprint arXiv:1703.08135, 2017.

[7] Kamper, H., Jansen, A. and Goldwater, S., "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *24*(4), pp.669-679, 2016.

[8] Siu, M.H., Gish, H., Lowe, S. and Chan, A., "Unsupervised audio patterns discovery using HMM-based self-organized units," *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

[9] Lööf, J., Gollan, C. and Ney, H., "Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a Polish speech recognition system," *Tenth Annual Conference of the International Speech Communication Association,* 2009.

[10] Jansen, A., Church, K. and Hermansky, H., "Towards spoken term discovery at scale with zero resources," *Eleventh Annual Conference of the International Speech Communication Association,* 2010.

[11] Ma, J., Matsoukas, S., Kimball, O. and Schwartz, R., "Unsupervised training on large amounts of broadcast news data," *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on* (Vol. 3, pp. III-III). IEEE, May 2006.

[12] Gish, H., Siu, M.H., Chan, A. and Belfield, B., "Unsupervised training of an HMM-based speech recognizer for topic classification," *Tenth Annual Conference of the International Speech Communication Association,* 2009.

[13] Chang-Ho Han, Choon-Suk Oh. "Implementation of a 3D Recognition applying Depth map and HMM," The Journal of The Institute of Webcasting, Internet and Telecommunication VOL. 12 No. 2, pp.119-126, December 2012.

[14] Sun-Jin Oh, "Design and Evaluation of a Weighted Intrusion Detection Method," The Journal of The Institute of Webcasting, Internet and Telecommunication VOL. 11 No. 3, pp. 181-188, June 2011.