

Hearing aid classification method based on improved AP algorithm

Chen Xiaomei^{1,a}, Ren Meina¹ and Zhong Bo²

¹ EDP Sciences, north china electric power university, school of Electrical and Electronic Engineering, China

² EDP Sciences, national institute of metrology, Acoustics group, China

Abstract. Based on the large medical data to evaluate the performance of the hearing aid is a promising way. Achieving the classification of the hearing aid is the foundation. In this paper an improved semi-supervised AP clustering algorithm based on density path is proposed. The PESQ score is taken as the substitution of subjective score for the speech segments, which is also taken as a semi-supervised basis to improve classification accuracy. The Euclidean distance similarity is improved based on the density path, making it suitable for complex shape data sets. Through experimental verification, compared with the traditional AP algorithm, the improved algorithm shows obvious advantages in terms of hearing aid classification accuracy and recognition performance.

1 Introduction

Hearing aids worn by different hearing patients are also different in style, and there are also great differences in their quality performance. The quality of hearing aids is of great importance to hearing impaired patients. A large amount of data was generated during the treatment process. How to effectively extract some valuable information from these data, distinguish hearing aids of different styles, and improve the detection level of hearing aid product performance have important theoretical and practical significance.

Clustering is a widely used method in data mining. At present, the AP selection algorithm is promising. it does not need to select the initial value, and allows the distribution of data in a non-Euler also allows unconventional point - point measurement method. But it also has shortcomings. Its complexity is high, and there are limitations to large-scale data clustering. literature [1] proposed an incremental AP algorithm which could improve the effectiveness when applied to large data, the literature [2] proposed the M-AP clustering algorithm and added a merge process to effectively solve complex shape dataset problems; Literature [3] summarized some methods and compared them, and determined that the performance of IGP indicators to find the best clustering number. But facing the complexity of hearing aids medical data, all the above AP related algorithms can't obtain acceptable results. In this paper, an improved AP algorithm is put forward and its effectiveness is evaluated.

2 Affinity Propagation Clustering Algorithm

AP clustering is a new clustering algorithm. Frey et al. [4] first proposed AP algorithm in Science in 2007 which is an unsupervised clustering algorithm. No need to determine the number of clusters and the cluster center in advance, instead, it uses all data points as potential cluster centers and clusters according to the similarity between data points [5]. Moreover, it does not need to select the initial clustering center, which effectively avoids the problem of the initial clustering center and improves the reliability of the clustering result.

Clustering is based on the sample's closeness in nature. In order to make the class reasonable, the similarity, that is the degree of distance between the samples, is described [6]. The AP clustering algorithm uses the Euclidean distance function as a measure of similarity. That is, the similarity s between any two points X_i and X_j is:

$$s(i, j) = -\|x_i - x_j\|^2, i \neq j \quad (1)$$

The AP algorithm passes two types of messages, Responsibility matrix r and Availabilities matrix a , by continually iteratively updating the two matrices until a stable number of high quality cluster centers are produced. The relationship between responsibility and availabilities is shown in the Fig. 1 below:

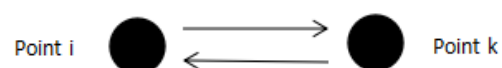


Fig. 1. Transferring messages between data points.

Where $r(i, k)$ represents the numerical message sent from point i to the candidate cluster center k , reflecting whether the k -point is suitable as the cluster

^a Corresponding author: zhongbo@nim.ac.cn.

center of the i -point. $a(i,k)$ is a numerical message sent from the candidate cluster center k to i , reflecting whether or not i has chosen k as its clustering center. The stronger $r(i,k)$ and $a(i,k)$, the greater the probability of k -point as the cluster center, and the more likely the i -point belongs to the cluster centered at the k -point cluster.

The iterative update formula is as follows:

$$r(i,k) = s(i,k) - \max_{k' \neq k} (a(i,k') + s(i,k')) \quad (2)$$

$$a(i,k) = \begin{cases} \min\{0, r(k,k) + \sum_{i' \in \{i,k\}} \max(0, r(i',k))\} & i \neq k \\ \sum_{i' \in \{i,k\}} \max(0, r(i',k)) & i = k \end{cases} \quad (3)$$

After each update, it can determine the representative sample point k of the current sample i , k is the k that gets the maximum value. If $i = k$, then it means that the sample i is the class representative point of its own cluster, if not, then explain that i belongs to the cluster to which k belongs.

3 Semi-supervised AP algorithm based on density path

3.1. Sample constraint information

The main idea of semi-supervised clustering algorithm is that using the known class of samples to form equivalent pairs of point constraint information, adjusting the similarity of the algorithm. Sample constraint information is divided into two kinds: Must-link, That is, two samples X_i and X_j must belong to the same class, expressed as $(X_i, X_j) \in Must-link$; Cannot-link, The restriction specifies that two samples X_i and X_j do not belong to the same category, expressed as $(X_i, X_j) \in Cannot-link$.

The similarity matrix was adjusted as following.

$$s(i,j) = \begin{cases} 0, (X_i, X_j) \in Mustlink \\ -Inf, (X_i, X_j) \in Cannotlink \end{cases} \quad (4)$$

3.2. Similarity measure based on density path

In general, the similarity of the nearest neighbor propagation algorithm uses Euclidean distance. The Euclidean distance of any two points is defined

as: $d(i,j) = \sqrt{\|X_i - X_j\|^2}$. Clustering is to divide data sample

points into multiple classes. Samples of the same class have high similarity, the Euclidean distance is small, while the sample similarity in different classes is low, and the Euclidean distance is relatively large. However, sometimes there will be situations as shown in Fig.2, $d(1,3) > d(1,2)$. According to the definition of clustering, sample 1 and sample 2 should be of the same type. The actual situation is that sample 1 and sample 3 are the same type, resulting in a contradiction. The reason is that the sample density is not considered when

clustering. In order to solve this problem, the literature [7] proposed the similarity of ε -Nearest Neighbor Distance. The purpose is to amplify the sample distance in the low density region and reduce the sample distance in the high density region.

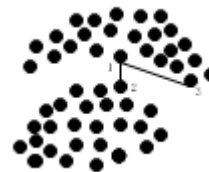


Fig. 2. Comparison of data sample point similarity

The ε -neighbor distance is defined as follows: in the sample space, construct an undirected weighted graph $G = (V, E)$, where V is the set of vertices, each vertex corresponds to a sample, and E is an edge set, which refers to the distance between each vertex. For any sample, its ε -neighbor distance.

$$G_\varepsilon(i,j) = \begin{cases} \varphi_1^{d(i,j)} - 1, d(i,j) \leq \varepsilon \\ \varphi_2^{d(i,j)} - 1, d(i,j) > \varepsilon \end{cases} \quad (5)$$

In the formula, φ_1 and φ_2 are the density adjustment factor. When finding the distance between the sample points, order $\varphi_1 < 1$, on the contrary, for distance, order $\varphi_2 > 1$; to ensure $d(i,j) = 0$, $G_\varepsilon(i,j) = 0$, subtract the constant 1. However, the ε -neighbor distance represents the degree of similarity, which may cause the points in the same density region to be stretched, that is, the samples i, j, k belong to the same class, and $d(i,j) \leq \varepsilon$, $d(j,k) \leq \varepsilon$, $d(i,k) > \varepsilon$. Using formula(5) to calculate the distance between sample i and sample k is stretched, it is possible to make $d(i,j) + d(j,k) \leq d(i,k)$, Causes sample i and sample k to belong to different classes, In order to solve this problem, the literature [7] proposed the definition of popular similarity, but this formula is too complicated to implement, and it takes a long time for large sample size data, in order to correctly reflect the similarity between sample data points, this article uses a density path based similarity measure.

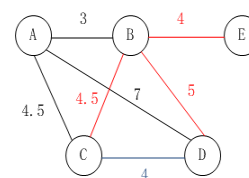


Fig. 3. Undirected weighted figure

In the undirected weighted graph, the adjusted ε -neighbor distance is used instead of the previous Euclidean distance. Definition: if there is a path between two sample points i and j , it means all paths from sample point i to sample point j . For each path $G_\varepsilon(i,j) \in G_\varepsilon$, and the path length between them is the weight value $G_\varepsilon(i,j)$; If there is no path between two vertices, their path length is infinite Inf. As shown in Fig.

3, there are multiple paths from one vertex to another. We call the path with the smallest weight as the shortest path and the length of the shortest path as the shortest distance [8].

The most commonly used solution to the shortest path problem is Dijkstra's algorithm [9]. The single-step search method is used to find the shortest path. We use this algorithm to find the measurement function based on the density path based on the ε -nearest neighbor distance:

$$D_{ij} = \min_{G_c(i,k) \in G_c} \left\{ \sum_{k=i}^j G_c(i,k) \right\} \quad (6)$$

Which D_{ij} satisfies nonnegative: if $D_{ij} \geq 0$, and only if $i = j$, $D_{ij} = 0$, it also satisfies the symmetry: $D_{ij} = D_{ji}$, triangle inequality: $D_{ij} \leq D_{ik} + D_{kj}$. The similarity matrix is replaced by a metric function. Since it is obtained from the shortest path, it can be connected with samples in the same high-density area with many shorter sides, and connect different density sample points with the longer side passing through the low-density area. In this way, it is possible to classify the sample points in the same density area as close to each other as possible, and sample points in different density areas are classified as different types. Effectively solve the problems raised before.

4 Experiment and result analysis

The data set is divided into two groups: the training data set and the test data set. The classifier is first constructed with training datasets, and then the validity of the constructed classifiers is tested with the detection dataset. Then the traditional AP algorithm is compared with the improved algorithm.

4.1 Algorithm validity indicator

In order to effectively evaluate the performance of clustering, the experiment will use the following three indicators:

(1) Classification accuracy (CR)

The classification accuracy (CR) represents the ratio of correctly classified samples to the total data set [10], which is defined as follows:

$$CR = \sum_{c=1}^k \frac{n_{\hat{c}}}{n_c} \times 100\% \quad (7)$$

Among them, n_c represents the number of samples contained in a class, $n_{\hat{c}}$ indicating the correct number of samples in this class, $c \{c=1,2,\dots,k\}$ is an arbitrary cluster. The higher the (CR), the more accurate the clustering and the higher the clustering accuracy [11].

(2) Square error sum

The squared sum of errors is often used as an objective function to construct the classifier and is often used to represent the classifier's distortion or coherence [12]. The square sum of errors is equal to the sum of the

squares of the distances from all samples to their class representative points:

$$SS = \sum_{c=1}^k \sum_{X_i \in C_c} dista(X_i, \gamma_k) \quad (8)$$

(3) F-measure indicator

The F-measure index is a commonly used indicator for clustering algorithm evaluation. It combines the "accuracy" and "recall" of the algorithm to measure the effectiveness of clustering [13]. For an arbitrary cluster, $C_i \in C = \{C_1, \dots, C_k\}$, the accuracy rate is $Pr ec_i = N_{\hat{i}} / N_i$, $N_{\hat{i}}$ is the correct number of samples in the C_i classification, N_i is the total number of samples in the C_i classification after classification; recall is defined as: $Re c_i = N_{\hat{i}} / N_{C_i}$, the N_{C_i} is the total number of samples C_i in the category without classification.. So the F-measure indicator is defined as: $F-measure(i) = 2P_i R_i / (P_i + R_i)$, then average it, $F-measure = \frac{1}{k} \sum_{i=1}^k F-measure(i)$. Its value range is [0, 1], the larger the value, the more accurate the algorithm [14].

4.2 Experiment results analysis

This article builds an experimental system and uses artificial ears to wear three different styles of hearing aids, namely A, B, and C, and then use the software to give it artificially eight different noises. They are 16000Hz babble, fl6, factory, leopard, m109, Motorcycle, pink, volvo in the standard noise speech library. The noise level is 30dB, The hearing aid selects one of two modes, namely mode 1 and mode 4, voice selects the speech 1 of the standard speech library, and uses software to divide it into 3-8s speech segments, removes the pauses in the speech, improves the detection quality, and clusters quality.

4.3 Analysis of results

The data during the experiment is 30dB. Since we are hearing aids in 3 different styles, we specify the number of clustering classes as 3, and the test results are analyzed according to different groups:

Mapping the clustering results in two-dimensional space and displaying them according to different clustering results. The horizontal axis represents different samples, and the vertical axis represents the corresponding PESQ score of each sample, as shown in the following figure:

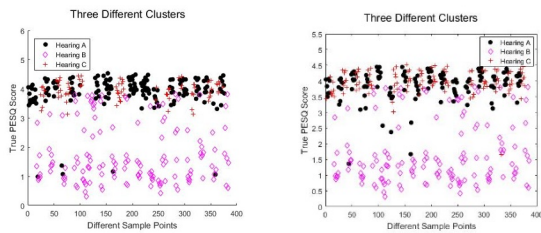
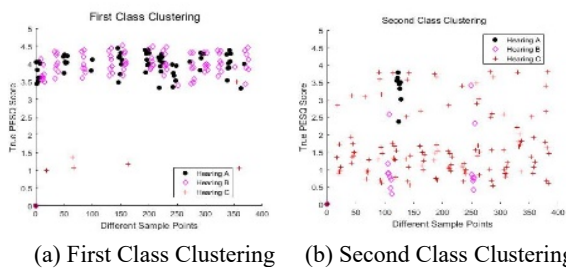
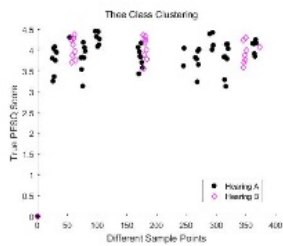


Fig. 4. Standard AP clustering

From Fig. 4 and Fig. 5, it can be seen that the samples are divided into 3 categories and the PESQ scores of different classes are not the same. Then the relationship between the different styles of hearing aids and PESQ corresponding to different classes is analyzed in detail. For the sample, the vertical axis is the PESQ score. Different colors represent different styles of hearing aids, as shown in the following figure.



(a) First Class Clustering (b) Second Class Clustering

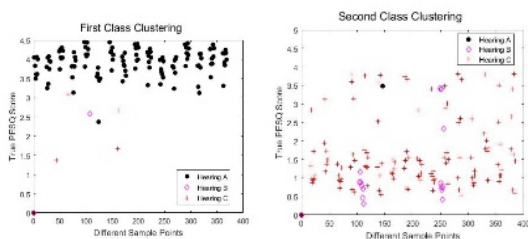


(c) Three Class Clustering

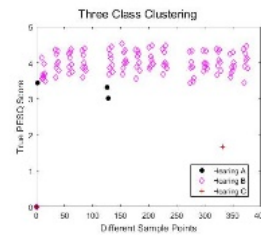
Fig. 6. Standard AP algorithm clustering different styles of hearing aid scatter plot

From (a) in Fig. 6, it can be seen that most of the cluster 1 samples are hearing aids A and B, and their PESQ scores are all above 3 points, and there are relatively more hearing aids B; From (b), most of the points in class 2 are hearing aids C, and most scores are below 2 points. In the third chart (c), most points in cluster 3 are hearing aids A and B, and most scores are all above 3 points. Among them, the style classification of the first and third types of hearing aids is not very accurate.

The improved AP algorithm results is shown in Fig. 7.



(a) First Cluster (b) Second Cluster



(c) Third Cluster

Fig. 7. Improved AP algorithm clustering different styles of hearing aid scatter plot

In (a) of Fig. 7, it can be seen that in most of the cluster 1 sample points are Hearing A, and their PESQ scores are all above 2.5 points; in the second graph (b), most of the cluster 2 points are Hearing Aid C. And most of the scores are below 2 points. In the third graph (c), the points of cluster 3 are all hearing aids B, and the scores are all above 2.5 points. The three clusters can represent three hearing aids of different styles. The comparison results show that the improved AP algorithm can accurately distinguish different styles of hearing aids, and the hearing aids A and B voice quality is significantly better than the C section.

The three evaluation indicators of clustering are shown in Table 1 below:

Table 1 three indicator values

Different indicators	Standard AP	Improved AP
Classification accuracy(CR)(%)	70.05	93.49
Square error sum	15.1866	14.2059
F-measure	0.6902	0.9348

According to the Table 1, the accuracy of the AP algorithm is higher than the standard AP; the squared error of the second indicator indicates the degree of distortion or cohesion of the cluster. The smaller the value, the smaller the distortion. Improved algorithm this indicator is relatively small, which proves that the distortion is small and the algorithm is better. The bigger the third indicator is, the more accurate it is. After the comparison, it shows that the improved algorithm is more accurate.

In summary, in the case of a noise level of 30 dB, the improved AP algorithm is superior to the conventional AP algorithm. This algorithm can distinguish hearing aids of different styles and can prove which hearing aid has better speech quality. Therefore, Good results have been achieved in the classification of hearing aids.

5 Conclusion

Experiments show that clustering is better applied to the speech quality classification of hearing aids, and the improved algorithm proposed in this paper is obviously better than the traditional AP algorithm. Its classification accuracy is high, and the squared error is small. F-measure value Large, and the classification can represent hearing aids of different styles, indirectly confirming that A and B hearing aids have better speech quality than C. Comprehensively, the changes in the indicators in the

three tables can also be found that the noise has an impact on the hearing aid's voice quality, so it also has a certain impact on the classification accuracy. The main work in the future is how to improve the accuracy of clustering under higher noise levels.

References

1. L.L.Sun.AP clustering algorithm research and its application in electronic medical record mining[D].Dalian:Dalian University of Technology, 2017: 101-110
2. Y.S.Gan.An improved algorithm:M-AP clustering algorithm[J].Computer Science,2015,42(1):232-235
3. S.B.Zhou.Comparative research on the method of determining the optimal number of clusters based on the nearest neighborpropagation algorithm[J]. Computer Science,2011,38(2):225-228
4. Frey B J, Dueck D.Clustering by passing messages between data points.Science,2007,315(5814): 972-976
5. H.X.Pan.A Modified Adaptive Filter[J].Journal of Artillery Launching and Control,2011
6. J.Sun. Application of similarity measure in gene expression cluster analysis[J].Modern Electronic Technology.2012
7. Z.Zhang.Semi-supervised Traffic Classification Based on Near-Adjacency Spread Learning[J].Acta Automatica Sinica,2013,37(7):1101-1109
8. Wagstaf K,Cardie C.Clustering with instance-level constraints.In: Proceedings of the 17th International Conference on Machine Learning, Stanford,USA:MorganKaufmann Publishers,2000.1103-1110
9. Y.L.Liao.Research on nearest neighbor propagation clustering algorithm and its application in high dimensional data[D].Guang zhou:South China University of Technology,2012.23-32
10. C.D.Wang,J.H.Lai,C.Suen,et al.Multi-Exemplar Affinity Propagation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35 (9):2223-2237.
11. D.W.Chen,J.Q.Sheng,J.J.Chen,et al. Stability-Based Preference Selection in Affinity Propagation[J].NeuralComputing&Applications,2014,25(7-8):1809-1822.
12. P.Li.Improvement of Affinity Propagation Clustering Algorithm and Its Application[D].Zhejiang:Zhejiang University,2017.50-51
13. C.H.Huang,J.Yin,Y.Hou. A text similarity measure method based on the TF-IDF method of term terms[J]. Chinese Journal of Computers, 2011, 34 (5):856-864
14. C.H.Yang.Application of AP algorithm in image Clustering [J].Computer and Digital Engineering, 2012,40(10):119-121