IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 5, Ver. V (Sep. – Oct. 2015), PP 26-29
www.iosrjournals.org

Music Data Organization Using Heuristic Hierarchical Agglomerative Co-Clustering

N. V. Keerthana[1], S. P. Yazhini[2]
[1]*Department of information Technology,Kongu Engineering college/Anna University,India*
[2]*Department of information Technology,Kongu Engineering college/Anna University,India*

Abstract: *In Data mining, Information Retrieval (IR) is the study of significant recent interest. In IR the Music Information Retrieval (MIR) is unitary of the challenging problems. At MIR, the utilization of Tags, Styles and Mood labels (T/S/M) can be extracted from some music websites. A typical problem is how to understand the relationship between the T/S/M. Co-clustering is the combination of two different types of data simultaneously. Hierarchical Co-clustering (HCC) extracts the acoustic features of music to find the similarity among the artist. In this paper, we systematically analyze the application of Heuristic Hierarchical Agglomerative Co-clustering (HHACC) method for music data organization. We demonstrate that this HHACC method achieves better performance than other clustering methods.*

Keywords: *Hierarchical Co-clustering, HACC, HHACC, T/S/M.*

I. Introduction

The music information retrieval research is the interdisciplinary science of retrieving information from music. It has a number of applications concerned with classification, clustering, indexing and searching in musical database. Traditional musical classification approaches usually assume the each piece of music has a unique style and they make use of the music content to construct the classifier. This classifier is identically used for classifying each piece of music information into it unique style. The music information is particularly differentiated as Tags, Style and Mood labels. The style and mood labels provide special features to define the similarity between the artist and the music tags. Generally the tags and labels are assigned to individual pitch, not to the artist. So by sampling the pitch of an artist, one is able to know tag, style and mood labels of the artist.

Hierarchical clustering is the method of analyzing the cluster, which is used to build a hierarchy of clusters. In hierarchical clustering the series of data partitions takes place, which may range from a single cluster containing all objects to the number of clusters each containing an exclusive target. Hierarchical clustering offers an extended description of document browsing [2]. There are two types of hierarchical clustering, one is divisive method and the other is agglomerative method. The former method divides the data set into smaller groups iteratively and the later one is the reverse process of divisive approach. Co-clustering or two-mode clustering is a data mining technique clustering of multiple data simultaneously. After analyzing with the proposed hierarchical divisive co-clustering (HDCC) method and hierarchical agglomerative co-clustering (HACC), we present a fictitious method, heuristic hierarchical agglomerative co-clustering (HACC) method. The divisive HDCC combines K-means and similar value decomposition (SVD) [1]. The HACC method starts with a single cluster and then iteratively merges two nearby clusters into one cluster until all the points are merged into a single cluster. In the case of HHACC, at each step of merging procedure, HHACC can merge a subset of the T/S/M labels and the subset of the artist. Thereby it needs to construct double-hierarchical for both artist and T/S/M. HACC merges the artist and T/S/M into a single group at the earliest possible stage [4].

Our finish is that search clusters with two types of data will be employed for better retrieval when both types of data designated in a query, e.g., given a query with an artist and one of its T/S/M, one can probably retrieve them together from a creative person-tag hierarchy, while with the query composed of an artist and style, one can retrieve simultaneously from an artist-style hierarchy. In this paper, we demonstrate that such mixed-data-type hierarchical cluster can generate HCC and empirically better cluster generated by concurrent usage of two data types.

Our contributions in this paper are: 1) we develop a fictional hierarchical agglomerative co-clustering method to organize a music data. 2) We analyze a demonstration to show that HHACC have the capacity of providing reasonable artist similarity qualification measures.

II. Existing Work

Hierarchical clustering is the operation of generating the clusters that take in a predetermine ordering in the form of tree like cluster structure of partitions. Hierarchical clustering algorithms organize input data either bottom up (agglomerative) or top down (divisive) [3]. Generally, hierarchical agglomerative clustering is more practiced than the hierarchical divisive clustering. Co-clustering is the clustering of more than one data

type. Dhillon [5] suggests the idea of modeling the document collection as a spectral bipartite graph between documents and words. J. Li *et al* [1] suggests two types of hierarchical co-clustering methods to organize the music data.

While hierarchical co-clustering deals with the constructing hierarchical structure for two or more data types, it attempts to achieve the purpose of both hierarchical clustering and co-clustering. Xu *et al* [6] suggested a hierarchical divisive co-clustering method to find out document clusters and associative cluster simultaneously. Though this hierarchical divisive co-clustering algorithm is proposed to our knowledge, few scholars have proposed the hierarchical agglomerative co-clustering (e.g., Li *et a*l [1] proposed a novel hierarchical agglomerative co-clustering method).

In late years, much research carried out **Constrained clustering-** integrating various kinds of background knowledge in the clustering operation. Existing constrained clustering methods have been focused in employment of background information in the sort of instance level "must-link" and "cannot-link" constraint, which, as the designation suggests, assert that, for a pair of data instance, they must be in the same cluster and they should be in distinct clusters, respectively. Latterly, there do exist a few works on incorporating constraints into hierarchical clustering (e.g., by drawing out the partial known hierarchy with the constraint to a full hierarchy or by changing the order of cluster merging process) [8]. Nevertheless, these constrained clustering methods cannot be applied to our heuristic hierarchical agglomerative co-clustering method.

III. Heuristic Hierarchical Agglomerative Co-Clustering Method

We begin the segment by reporting the details of our application of heuristic hierarchical agglomerative co-clustering algorithm to the problem of co-clustering artist-T/S/M. This procedure is similar to that in Li *et al* [1]. We then deliver a novel heuristic hierarchical agglomerative co-clustering algorithm called HHACC, which could likewise be used to cluster artist-T/S/M.

1. Problem Definition

Given a set of k artist, $\boldsymbol{R}$**=** **{**r_1, r_2,..., r_k**}**, and a set of l T/S/M, $\boldsymbol{S}$**=** {s_1, s_2,…, s_l}. The artist-T/S/M relationship matrix as $k \times l$, $\boldsymbol{Y} = (y_{ij})$, such that y_{ij} represents the relationship between the ith artist in $\boldsymbol{R}$ and the jth T/S/M in $\boldsymbol{S}$. Our goal is to generate a heuristic optimized hierarchical clustering of $\boldsymbol{R}$ and $\boldsymbol{S}$ based on the matrix $\boldsymbol{Y}$ so that each artist-T/S/M can be in the corresponding cluster and show the hierarchical relationship between these clusters.

2. Heuristic Hierarchical Agglomerative Co-Clustering

Here we present our heuristic hierarchical agglomerative co-clustering (HHACC) algorithm. Like the traditional agglomerative hierarchical clustering algorithms, HACC starts with single cluster and successfully combines the two nearest clusters until one cluster is gone off. Yet unlike traditional algorithms, it may unite two different information types. The output of HHACC is thus a single tree where the leaves are rows and columns of the input matrix, where nodes having both rows and columns as descendants may appear in any non-leaf stage. The HHACC algorithm is presented in algorithm 1. The method TwoNodePick is for selecting two nodes to combine.

Algorithm 1: HHACC Algorithm
Given *R* -set of artists and *S*- set of T/S/Ms

Create an empty list H
L ← *attributes in R + attributes in S*
M ← *size [R] +size [S]*
Bottom layer ← *L+H*
For c=0 to *M* – 1 ***do***
i, j = TwoNodePick (L)
O= Merge (*i, j)*
End for
Apply heuristic
Remove *i, j from L* and do *O+ L*
Next level ← *I +H*

IV. Implementation

1. Data Set

A data set is a collection of data. Every column of the table represents a particular variable, and each row corresponds to a given number of the dataset. Each value is known as a datum. A data set consisting of 1330 songs and 70 artists is collected. For each artist, T/S/M is assigned. The feature extraction values are

obtained by extracting Mel-Frequency Cepstral Co-efficients (MFCC) and Short-Term Fourier Features (STFF) values from JAudio tool. Note that an artist may receive the same tag more than once, while being assigned the same style/mood only once. Table 1 contains sample T/S/M.

Tags	**Styles**	**Mood Labels**
Oscar	Hip-hop	Running
60s	Indie jazz	Joy
Shore	Cape jazz	Dreamy
80s	Neo soul	Sad
21st	Folk	Energetic
Ceremony	Classic	Romance
Warm	Provocative	Breezing
	Soft rock	Happy
	Duet	Driving

Table 1. List of T/S/M

2. Hierarchy Generation

Hierarchy generation is an important step in the cluster generation process. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training is difficult. Hierarchy generation steps take a considerable amount of processing time.

3. Feature Extraction

Feature extraction is the procedure of generating features to be used in the selection and classification tasks. There has been a considerable amount of work in extracting the descriptive features of the music for music genre classification and artist identification. The purpose of descriptive feature for a specific application is the main challenge in building pattern recognition schemes. In one case the feature is extracted standard machine learning techniques independent of the specific application area can be habituated. The feature set consists of two processing methods such as, Speech Processing and Timbral Feature Extraction.

3.1. Speech Processing

MFCC is the feature set that is extremely popular in speedy processing. It is planned to capture short-term spectral based feature. The characteristics are computed as follows: Initially, the logarithm of the amplitude spectrum based on short-term Fourier transform is computed, where the frequencies are divided into 13 bins using the Mel-Frequency scale [9].

3.2. Timbral Feature Extraction

STFT is a set of feature associated with the temporal texture and it is not analyzed using MFCC. It consists of five types: spectral Centroid, Spectral Rolloff, Spectral Flux, Zero Crossings and Low Energy. More elaborate description of STFT can be found in [5].

4. Music Data Organization

The organization of data is important to the clustering system. The performance objective of the clustering system must be achieved while, regardless of the actual physical database structure, allowing the application to access data as simple, logical structures. With the extracted feature of each song, the inter cluster distance between the acoustic features that represent the songs of different artist is evaluated. Finally the average of all the inter cluster distances between the artist-T/S/M relationship is evaluated. From the results, it needs to observe that the quantified artist similarity match very closely to artist similarity based on acoustic features of their music recordings. The average distance increase one by one the combined similarity decreases almost constantly.

The inter cluster distance is the distance between two points that one measure with a ruler. In one dimension, the distance between two points on the real line is the absolute value of their numerical distance. Thus, if x and y are the two points on the real line. Then the distance between them is given by the Equation (1).

$$D(x_i, y_j) = \sqrt{(x_i - y_j)^2} = |x_i - y_j| \quad \ldots\ldots (1)$$

V. Performance Analysis

The experimental setup consisted of a data set of the music data. The music data are classified into: artists, T/S/M, MFCC value and STFT. This clearly demonstrates the effects of using the concepts in the

HHACC process. To evaluate the artist-T/S/M relationships, we utilize CoPhenetic Correlation Coefficient (CPCC) [11] as an evaluation measure. CPCC is given in the Equation (2) [1].

$$\text{CPCC} = \frac{\sum_{i<j}\left(d(x_i,y_j)-d\right)\left(t(x_i,y_j)-t\right)}{\sqrt{\left(\sum_{i<j}\left(d(x_i,y_j)-d\right)^2\right)\left(\sum_{i<j}\left(t(x_i,y_j)-t\right)^2\right)}} \quad \ldots\ldots (2)$$

Here, $d(x_i, y_j)$ and $t(x_i, y_j)$ are the inter cluster distance and dendrogrammatic distance between the *i*th and *j*th data points. The comparison results based on CPCC are shown in Fig. 1, and so that the HHACC can generate a good relationship of artist-T/S/M than HACC [1].

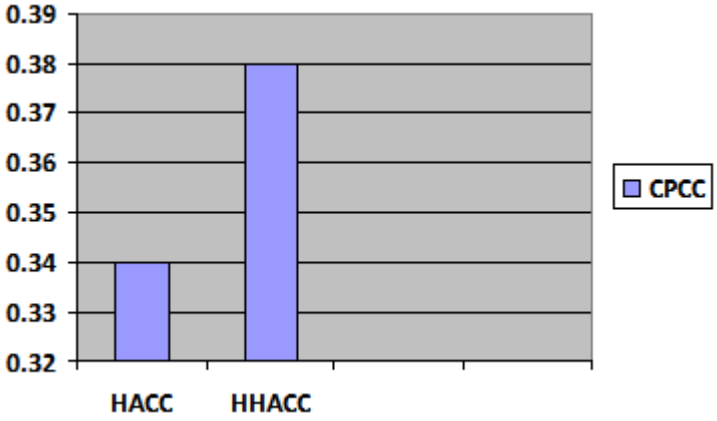

Figure 1. CPCC of HACC and HHACC

VI. Conclusion

In this paper, we systematically study the usage of heuristic hierarchical agglomerative co-clustering methods for organizing the music data. In particular our proposed HHACC method outperforms the other algorithms. Moreover the HHACC algorithm with the complete performance is used in better understanding of the relationship between artist and artist related information. There are several avenues for future research. First, we will investigate the HHACC algorithm will incorporate layer-wise optimization other than cluster heterogeneity for cluster merging process and to implement HHACC method to organize video data streams.

References

[1]. Jingxuan Li, Bo Shao, Tao Li, and Mitsunori Ogihara, "Hierarchical Co-Clustering: A New Way to Organize the Music Data" *IEEE Transactions On Multimedia*, vol. 14, no. 2, pp. 471-481, 2012.

[2]. D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, "Scatter/gather: A Cluster-based approach to browsing large document collection," in *Proc. 15th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 1992, pp. 318-329.

[3]. P. N. Tan et al., "Introduction to Data Mining," Boston, *MA: Pearson Addison Wesley*, 2006.

[4]. J. Li, T. Li, and M. Ogihara, "Hierarchical co-clustering of artists and tags," in *Proc. 11th Int. Society for Music Information Retrieval Conf.* (ISMR 2010).

[5]. I. S. Dhillon, "Co-Clustering document and words using bipartite spectral graph partitioning," in *Proc. 7th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2001, pp. 269-274.

[6]. G. Xu and W. Y. Ma, "Building implicit links from content for forum search," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Research and Devolopement in Infrmation Retrieval (SIGIR'06)*, New York, 2006, pp. 300-307.

[7]. Rahmat Widia Sembiring, Jasni Mohamad Zain, Abdullah Embong, "A Comparative Agglomerative Hierarchical Clustering Method to Cluster Implemented Course," *Journal Of Computing, ISSN 2151-9617*, vol 2, Issue 12, December 2010.

[8]. K. Bade and A Nurnberger, "Creating a cluster hierarchy under constraints of a partially known hierarchy," in *Proc. 2008 SIAM Int. Conf. Data Mining*, 2008, pp. 13-24.

[9]. B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," in Proc. *Int. Symp. Music Information Research (ISMIR)*, 2000.

[10]. Y. Haixia, Z. Qiang, H. Xianlong, L. Zhuoyuan, "Efficient Hierarchical Algorithm For Mixed Mode Placement In Three Dimensional Integrated Circuit Chip Designs," *Tsinghua Science And Technology ISSN 1007-0214 03/18*, pp.161-169 vol 14, no 2, 2009.

[11]. R. R. Sokal and F. J. Rohlf, "The comparison of dendrograms by objective methods," Taxon, vol. 11, no. 2, pp. 33-40, 1962.

[12]. Dr. N. Rajalingam and K. Ranjini, "Hierarchical Clustering Algorithm - A Comparative Study," *International Journal of Computer Applications* , vol 19, no.3, April 2011.

[13]. D. Hui, Z. Qiang and B. Jinian, "Markov Clustering-Based Placement Algorithm for Hierarchical FPGAs," *Tsinghua Science and Technology, ISSN 1007-0214 10/17,* vol 16, no 1, pp62-68, February 2011.

[14]. B. Logan and A.Salomon, "A Content-Based Music Similarity Function," *Cambridge Research Labs-Tech Report*, 2001.

[15]. I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic Co-clustering," in *Proc. 9th ACM SIGKDD Int.Conf.Knowledge Discovery and Data Mining*, 2003, pp. 89-98.