CLUSTER AWARE NORMALIZATION FOR ENHANCING AUDIO SIMILARITY

Mathieu Lagrange[⋆] *Luis Gustavo Martins*[‡] *George Tzanetakis*[†]

[⋆] IRCAM CNRS, 1, place Igor Stravinsky, 75004 Paris, France, lagrange@ircam.fr
[‡] Portuguese Catholic University - School of Arts / CITAR, Porto, Portugal, lmartins@porto.ucp.pt
[†] Department of Computer Science, University of Victoria, BC, Canada, gtzan@uvic.ca

ABSTRACT

An important task in Music Information Retrieval is content-based similarity retrieval in which given a query music track, a set of tracks that are similar in terms of musical content are retrieved. A variety of audio features that attempt to model different aspects of the music have been proposed. In most cases the resulting audio feature vector used to represent each music track is high dimensional. It has been observed that high dimensional music similarity spaces exhibit some anomalies: hubs which are tracks that are similar to many other tracks, and orphans which are tracks that are not similar to most other tracks. These anomalies are an artifact of the high dimensional representation rather than actually based on the musical content. In this work we describe a distance normalization method that is shown to reduce the number of hubs and orphans. It is based on post-processing the similarity matrix that encodes the pair-wise track similarities and utilizes clustering to adapt the distance normalization to the local structure of the feature space.

Index Terms— distance normalization, information retrieval, kernel-based clustering

1. INTRODUCTION

Searching large databases for objects that have similar properties is a key task in many data mining applications. These objects are typically represented as a series of scalar features and therefore can be viewed as points in a vector space with dimensionality equal to the number of features. Using this representation the search for similar objects reduces to finding vectors that are neighbors of the feature vector representing the query object. In some cases, the only information at hand for achieving such a task is the distance between every pair of objects. Unfortunately, this raw information usually lacks consistency. As the dimensionality of the feature space gets higher, some objects get irrelevantly close or far from any other object [1, 2]. These objects are respectively coined "hubs" (which are irrelevantly close to many other objects) and "orphans" (which are irrelevantly far to many other objects). Hubs and orphans shall be avoided, as in practice they will respectively be often / never returned by ranking systems. Consequently, reducing them shall also lead to an improvement of precision and recall in ranking-based MIR applications.

Therefore, an alternative approach is to consider local normalization approaches which consider statistics that are computed over a given neighborhood [3]. The main issue is how to define K, i.e the size of the neighborhood to consider for the normalization. In [3], K is set to a unique arbitrarily small number with respect to the cardinality of the set of objects, N, and typically $K \in [5, 20] \ll N$. It is shown in [4] that this arbitrary setting is equivalent to the hypothesis that the set of objects is organized in clusters each of cardinality $\#C_l = K, \forall l$ (where l is the cluster label).

However, these clusters of objects (which correspond to actual classes) may be heterogeneous in size. We therefore propose to consider a normalization factor that is independently set for each object o_i, expressed as a function of the cardinality of the class it belongs to

$$K_i = \#C_l, o_i \in C_l. \quad (1)$$

The contribution of this paper is twofold. First, we demonstrate that, assuming prior knowledge of the organization of the data in classes/clusters of potentially heterogeneous cardinality, one can enhance several desirable properties of the affinity matrix. That is, the resulting cluster aware local normalization scheme is consistently better than other schemes in terms of accuracy and reversibility improvement. Secondly, we demonstrate that such normalization can be performed in an unsupervised manner, by estimating the cluster organization with the data at hand.

2. BACKGROUND

Let us consider N objects o_i, each described by a set of features f_i, upon which an distance matrix D encodes the pairwise distance d_{ij} between each each couple of objects, i and

This work has been supported by the "Agence Nationale de la Recherche" (French National Funding Agency) in the scope of the JCJC project HOULE and "Fundação para a Ciência e Tecnologia" (Portuguese National funds) in the scope of the project ref. PTDC/EIA-CCO/111050/2009.

j. It is proposed in [3] to infer from that distance an affinity matrix A_{ij} using a normalization factor specific to each object:

$$a_{ij} = \exp \frac{-d_{ij}^2}{\sigma_i \sigma_j} \tag{2}$$

where σ_i, σ_j are the scaling parameters for objects i and j respectively.

More precisely, a Local Scaling (LS) is performed in [3], by considering a neighborhood of a given object to normalize its affinity towards the other objects. In this case,

$$\sigma_i = d_{ib} \tag{3}$$

where b is the index of the K^{th} nearest neighbor of o_i.

Let us consider Equation 2 as the simpler function $a = e^{-d^*}$, where

$$d_{ij}^* = \frac{d_{ij}}{\sqrt{\sigma_i \sigma_j}} \tag{4}$$

Working directly on the distance matrix allow us to consider an iterative version of this normalization scheme by iteratively processing the resulting distance matrix several times.

3. DETERMINING K

Local schemes need K (the size of the neighborhood considered for normalization) to be set by the end user. In [3], K is set a priori for convenience to a small value. In the experiments reported in [5], the authors observed that, after a given value ($K = 25$), increasing K did not improve nor decreased significantly the accuracy. Even though such arbitrary setting may be convenient, it is may be desirable to set this value according to some statistics over the dataset at hand, as studied in [6]. It has been hypothesized that setting K to a given value is equivalent to assuming that the dataset is roughly organized as set of clusters with cardinality equal to K [6]. In fact, and as shown in [7], setting K as a low value is harmful as far as accuracy is concerned, and the maximal performance is reached when K is around the number of elements within each class. When dealing with realistic data, several phenomena can influence the optimal K setting. For example the presence of outliers supports considering a smaller K than the number of elements within each class.

4. PROPOSED APPROACH

We therefore propose to consider an alternate approach that roots the normalization process with perfect or estimated knowledge of the organization of the objects within the dataset.

4.1. Assuming knowledge of the class assignments

Instead of considering a fixed size of neighborhood to perform a local scaling of a given distance metric, we propose to consider a neighborhood size K specific for every object in a cluster. Brute force optimization of such sizes is impractical for datasets of reasonable size. So, in order to set this value, we propose to consider the cardinality of the cluster to which the object belongs:

$$K_i = \#C_l, o_i \in C_l \tag{5}$$

Thus, this approach is a supervised normalization scheme where the class labels of each object in the dataset are used to compute the normalization factors.

4.2. Estimating the cluster assignments using kernel based clustering

In realistic settings, knowledge of the class labels can not be assumed, whereas an approximate estimate of the number of classes can be guessed more easily by considering the number of abstract classes or concepts that the end use is interested in.

For that reason, we propose to estimate the cluster organization by means of a clustering step performed using the available features at hand. For the sake of simplicity, we considered a raw implementation of the kernel k-means algorithm [8]. However, depending on the specificity of the application scenario of interest, one may consider alternatives approaches. In large scale problems highly efficient approaches based on heuristics such as the ones used considered for community detection [9] can be also considered. Spectral clustering approaches can also be applied for enhanced clustering accuracy [10].

A qualitative analysis of the proposed approach can be taken by considering Figure 1. It depicts the distribution of distances between a given object and the others as vertical density histograms. The distributions of the raw distance matrix are not centered. Objects with low centered distribution are potential hubs, and object with high centered distribution are potential orphans. Normalizing using LS drastically changed the distributions, imposing a hard threshold for small distances. On contrary, the proposed approach preserves more diversity while minimizing the spread of the distributions.

5. EXPERIMENTS

In this section, we quantitatively compare the LS approach to the proposed scheme in order to give answers to the following questions:

1. Shall the normalizations be performed iteratively?
2. What is the gain of considering the cluster organization, be it known or estimated, when performing a local normalization?

Different acronyms for the methods under evaluation are used. The *i* prefix stands for the iterative LS version. For adaptive versions, the neighborhood size can be assumed as a

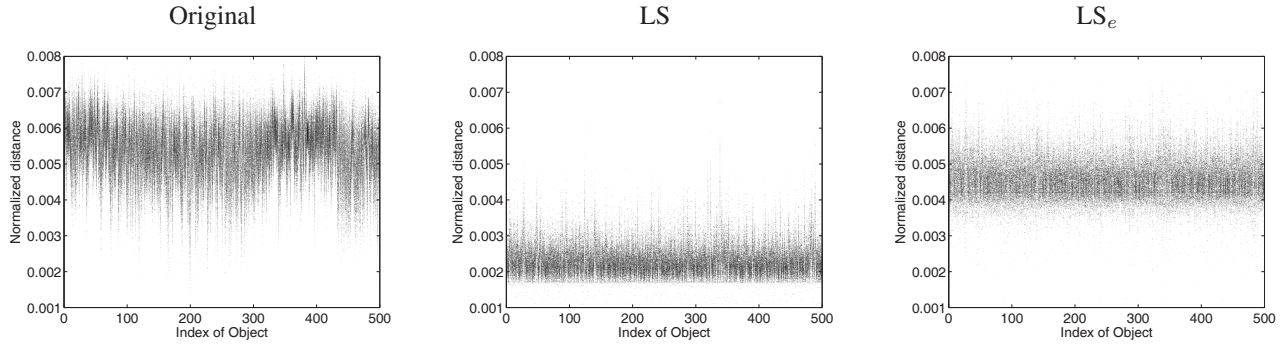

Fig. 1. *Distribution of distances between an object and the remaining objects as vertical density histograms (dark means high density) for a given distance matrix, unprocessed (left), processed with iterative LS (middle) or with the proposed approach (right).*

prior (LS$_p$) or estimated via clustering (LS$_e$). In the case of clustering, the considered clustering is the one that achieved the lowest average within-cluster distance over 50 runs of the kernel k-means with random initialization. Unless otherwise stated, the number of clusters is set according to the number of classes, *e.g.* it is considered as a prior.

For each normalization method, the method is either run once or iterated until convergence. The number of iterations is limited to 100, though convergence is achieved in less than 20 iterations in most cases. For the LS method, the neighborhood size K is set to 10.

5.1. Evaluation Datasets

A wide range of feature sets taken from three datasets of music and speech were considered (see Table 1).

The Cal500 dataset[1] comprises 502 songs. Among the five kernels considered in this study, three are computed from the audio and are supposed to reflect the timbre, rhythm and the harmony. The last two are respectively computed from social tags and the results of web search. More details can be found in [11].

The Magnatagatune dataset[2] comprises 5393 songs which have been organized into 229 clusters according to genre annotations given by listeners. The distance metric considered in this study was obtained by computing the pair-wise similarity between the songs using a state of the art content based approach. More details can be found in [7].

The Timit dataset is an acoustic phonetic speech corpus widely used in the speaker recognition community. In this study, we consider the sub corpus composed by Rob Tibshirani[3], that is made of 5 phonemes spoken by 50 different males speakers. The distance metric considered is a linear kernel of 4509 log periodograms, each representing 30 ms of speech.

For gaining statistical significance and speeding-up computation, each of those feature sets is randomly sampled to build 100 subsets of 500 randomly picked objects. The various normalization methods are applied to the resulting pairwise distance matrices of size 500 by 500 and the results are evaluated using the metrics described in the following section.

5.2. Evaluation Metrics

Retrieval of objects within a database that share properties with a given query is perhaps the simplest way of evaluating the effectiveness of a distance measure. In this paper, we consider the Mean Average Precision (MAP) as a measure of accuracy. The MAP is routinely employed in a wide variety of tasks in the information retrieval community [12].

As stated before, for a good accuracy, hubs and orphans shall be minimized. To account more precisely for such behaviors, we consider the so-called R-precision, which measures the ratio of the number of relevant documents to the number of retrieved documents, when all relevant documents have been retrieved (i.e. precision at recall equal to one). In that respect, it account well for such unwanted behaviors. Hubs are irrelevantly close neighbors that will therefore decrease the R-precision. Orphans are never close to any items. This will aversely decrease the R-precision.

5.3. Results

Cluster awareness

As expected, and as presented in Table 2, when the cluster information is known as a prior, the effectiveness of the normalization is high. This MAP gain decreases when considering the cluster information given by the clustering. Roughly, the accuracy gain of the methods based on the estimated cluster information is approximately half way in between the gains obtained by the constant setting and prior knowledge.

[1] http://cosmal.ucsd.edu/cal/projects/AnnRet/

[2] http://tagatune.org/Magnatagatune.html

[3] http://www-stat-class.stanford.edu/~tibs/ElemStatLearn/data.html

Name	Size	Nb Clus.	Size Clus.	Kernels
Cal500	502	20	25 (10)	6
Magna.	5393	229	24 (20)	1
Timit	4509	5	902 (192)	1

Table 1. *Synthetic overview of the different kernels considered. The cluster size is expressed in terms of average (first value) and standard deviation (second value).*

Concerning the R-precision, the cluster information is highly beneficial, be it given as a prior or estimated. This lead us to conclude that even if the cluster knowledge is not perfect, it is highly beneficial for better reducing hubs and orphans.

Iterative vs. non iterative normalization

For the Cal500 and Magnatagatune datasets, the use of the iterative version does not provide any gain (see Table 2). On the Timit dataset, the use of the iterative version is positive when taking into account the cluster structure of the data.

(a) MAP

	Ref	LS	iLS	LS_e	iLS_e	LS_p	iLS_p
Cal500	.121	122	.121	**.124**	**.124**	.126	.13
Magna.	.183	**.192**	.191	**.192**	.191	.193	.192
Timit	.725	.735	.731	.741	**.744**	.747	.752

(b) R-precision

	Ref	LS	iLS	LS_e	iLS_e	LS_p	iLS_p
Cal500	.134	.135	.135	.138	**.139**	.139	.142
Magna	.459	**.462**	.461	**.462**	.461	.463	.463
Timit	.686	.691	.689	.701	**.703**	.703	.706

Table 2. *Average MAP (a) and R-precision (b) for several databases before (Ref) and after non-iterative and iterative normalizations.*

6. CONCLUSION

A new contextual scaling approach has been introduced and tested over three datasets issued from the music and speech areas. Experiments showed that the proposed approach reduces unwanted phenomena like hubs and orphans, which results in an accuracy improvement when retrieving objects from a dataset.

Future work will focus on experimenting the proposed approach on datasets of different kind and geometry as well as evaluating the robustness of the approach while considering an imperfect prior knowledge of the number of clusters.

7. REFERENCES

[1] Milos Radovanovic, Alexandros Nanopoulos, and Ivanovic Mirjana, "Nearest Neighbors in High-Dimensional Data : The Emergence and Influence of Hubs," in *Proceedings of the 26th International Conference on Machine Learning*, 2009.

[2] Jean-Julien Aucouturier and F Pachet, "A scale-free distribution of false positives for a large class of audio similarity measures," *Pattern Recognition*, vol. 41, no. 1, pp. 272–284, Jan. 2008.

[3] L Zelnik-Manor and P Perona, "Self-Tuning Spectral Clustering," in *Annual Conference on Neural Information Processing Systems*, 2004.

[4] Mathieu Lagrange and George Tzanetakis, "Adaptive N-Normalization for enhancing Music Similarity," in *Proc. ICASSP*, 2011, pp. 1–4.

[5] Tim Pohle, Peter Knees, Markus Schedl, and Gerhard Widmer, "Automatically Adapting the Structure of Audio Similarity Spaces," in *Proceedings of the 1st Workshop on Learning the Semantics of Audio Signals*, 2006, pp. 66–75.

[6] Mathieu Lagrange and Joan Serra, "Unsupervised Accuracy improvment for Cover Song Detection using Spectral Connectivity Network," in *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, 2010, pp. 595–600.

[7] M. Lagrange and G. Tzanetakis, "Adaptive N-normalization for Enhancing Music Similarity," in *ICASSP 2011*, 2011, pp. 1–4.

[8] John Shawe-Taylor and Nello Cristianini, *Kernel methods for pattern analysis*, Cambridge University Press, 2004.

[9] VD Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, "Fast unfolding of communities in large networks," *Statistical Mechanics*, vol. 10, 2008.

[10] A. Ng and M. Jordan, "On spectral clustering: Analysis and an algorithm," in *Advance on Neural Information Processing Systems*, 2001.

[11] Luke Barrington, M. Yazdani, D. Turnbull, and Gert Lanckriet, "Combining feature kernels for semantic music retrieval," in *Proceedings of ISMIR*, 2008.

[12] CD Manning and P Raghavan, *Introduction to information retrieval*, Cambridge University Press, 2008.