Adaptive N-normalization for enhancing music similarity

Mathieu Lagrange, George Tzanetakis

ADAPTIVE N-NORMALIZATION FOR ENHANCING MUSIC SIMILARITY

Mathieu Lagrange

IRCAM CNRS
1, place Igor Stravinsky,
75004 PARIS - FRANCE
lagrange@ircam.fr

George Tzanetakis

Computer Science Department
University of Victoria,
BC, Canada
gtzan@uvic.ca

ABSTRACT

The N-Normalization is an efficient method for normalizing a given similarity computed among multimedia objects. It can be considered for clustering and kernel enhancement. However, most approaches to N-Normalization parametrize the method arbitrarily in an ad-hoc manner. In this paper, we show that the optimal parameterization is tightly related to the geometry of the problem at hand. For that purpose, we propose a method for estimating an optimal parameterization given only the associated pair-wise similarities computed from any specific dataset. This allows us to normalize the similarity in a meaningful manner. More specifically, the proposed method allows us to improve retrieval performance as well as minimize unwanted phenomena such as hubs and orphans.

Index Terms— Metric spaces, Normalization, Music Similarity

1. INTRODUCTION

Computing the distance or the similarity between some elements of interest is the first step in many tasks such as content-based retrieval, classification and clustering. Although each of those tasks have specific needs, one usually wants to ensure that the distance is such that: *"one item of a given class has its closest neighbors belonging to the same class"*.

Unfortunately, it has been shown that computing the similarity between complex elements described by noisy and high dimensional features usually leads to a distance metric plagued with many undesirable properties. Those observations are valid for the similarity amongst music segments [1] as well as many other tasks [2]. Some elements, the so-called "hubs", appear to be close to any other element while other elements, the so-called "orphans" are far from any other elements.

The N-normalization method has been shown to efficiently enhance the similarity metric [3], [4]. In these works, N is empirically set to a small value with respect to the dataset size, $N << S$. As we will demonstrate, in most settings, the optimal value strongly depends on the geometry of the dataset.

Therefore, there are two main contributions in this paper. First, we demonstrate that the optimal value of N is tightly linked to the geometry of the data set, more precisely the number of elements within each class and that parametrizing the normalization according to the data set is beneficial. Second, we introduce a method for estimating the parameter N using a statistical metric similar to the gap statistic proposed for detecting the number of clusters [5].

M.L. has been partially funded by the OSEO Quaero project.

2. BACKGROUND

Defining the similarity amongst a large number of elements is a fundamental problem in many information retrieval tasks. As far as music clips are concerned, the "bag-of-frames" approach is largely used where the audio signal is split into potentially overlapping frames. Each of those frames is modeled as a set of features accounting for the most important aspects of music, namely timbre, rhythm and harmony. A prototypical implementation is to model the frames of a given musical song using Gaussian Mixture Models (GMMs) of Mel-Frequency Cepstrum Components (MFCCs) [1]. A Query By Example (QBE) system built on this principle would compute for a given query its GMM model that would be compared to each model of the entry of the database using a given distance. Ranking those entries according to this distance then allows us to retrieve the "closest" songs to the query.

Although a lot can be done at the first steps, like providing a richer representation of the polyphony [6], using more diverse features [4], and considering different statistical models [1], we will focus in this paper on an efficient post-processing method that potentially improves the performance of the QBE by considering some statistics computed over the database.

For that purpose, if the accuracy of the QBE is high, one can consider the result of a clustering step in order to set to a high similarity the couple of elements that are identified as belonging to the same class [7]. If the accuracy of the QBE is low, one can consider spectral connectivity approaches as proposed in [8].

For large scale problems, one needs computationally simple methods such as the N-Normalization. This normalization have been used for identifying outliers [9], improving clustering [3], and more recently improving music similarity [4]. Within most of those approaches, the tuning parameter N is fixed a priori.

In this paper, unless stated otherwise, we use a reference QBE system and an evaluation database which are both publicly available and described in more detail in Section 6.

3. EVALUATION METRICS

3.1. Human and Automatic Evaluation of Retrieval Effectiveness

Ultimately the effectiveness of any query-by-example (QBE) system needs to be evaluated by humans. This is a time consuming process that is typically only conducted during large scale comparative evaluations of different systems. In the field of music information retrieval, the Music Information Retrieval Evaluation Exchange

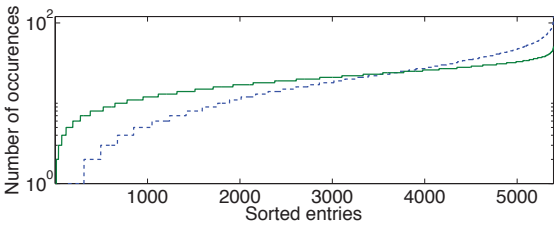

Fig. 1. *Sorted number of times an query appeared in a top 20 list of all the entries of the database before (solid line) and after 50-Normalization (dashed line).*

(MIREX) is an example of such a comparative evaluation. For example in MIREX 2010 audio-based music similarity and retrieval was evaluated using a data-set of 7000 clips (each 30 seconds long) from 10 genre groups. 100 songs (10 per genre group) were selected as queries and the 5 most similar songs to these queries according to each submitted algorithm were evaluated by the human graders. Songs by the same artist were omitted from the returned results. For each query/candidate pair the graders were asked to provide a broad score (not-similar, somewhat similar, very similar) and a fine score (a number between 0 and 100). These scores result in the Average Broad Score (ABS) and Average Fine Score (AFS) metrics.

In order to approximate this evaluation process using objective measures, one can consider the Artist-filtered Genre Precision (AGP), as it is nicely correlated with the subjective measures based on human evaluations over the last MIREX runs. As genre labels are frequently available for the clips of interest this measure can be calculated automatically. A good QBE system for a query of a given genre should return as closest elements mostly clips that belong to the same class. The k-AGP is defined as the the number of songs from the same genre as the query from the set of the k closest songs to the query excluding clips from the same artist. In this paper, we set K equal to 5 which is also the value used in MIREX.

3.2. Quantifying undesired properties

As shown in many studies [1] [2], undesired properties appear when dealing with elements compared within high dimensional vector spaces. These include the so-called "hubs" which are irrelevantly close to many other elements and the so-called "orphans", which are irrelevantly far to many other elements. In order to visualize such undesired properties one usually counts the number of time a given query is found in a top 20 list of every entries in the database. By sorting those counts, a curve such as the ones plotted on Figure 1 can be generated. On this figure, the dashed line depicts the counts obtained based on the results of the reference QBE. By considering the bottom left of the figure, one can see that orphans are present, since some entries are never close to any other elements. By considering the top right of the figure, one can see that hubs are also present, since some entries are close to many other elements. One can quantitatively measure orphans by considering the ratio r_o between the number of queries that are never in any top 20 lists versus the size of the database. For hubs, we count the maximum number of times a query was in a top 20 list, noted n_h or the ratio between n_h and the cardinality of the dataset. For the reference QBE and data-set used in the paper these values are $r_o = 0.025$ and $r_h = 0.0223$.

4. N-NORMALIZATION

Consider a square and symmetric matrix d that encodes the output of a given QBE system over a given data-set:

$$d(i,j) = \text{QBE}(i,j) \tag{1}$$

where each element of the matrix is the pairwise distance in the data-set. The N-normalized version of d is:

$$d_N(i,j) = \frac{d(i,j)}{\sqrt{d(i,i_N)d(j,j_N)}} \tag{2}$$

where i_N is the Nth neighbor of element i. This operation has been considered for enhancing spectral clustering under the term "local scaling" in [3], and for improving retrieval in musical databases [4]. Such normalization, or scaling, is valuable as it accounts for the distribution of neighbors of a given entry in order to weight its distance to other entries. For clustering, it allows us to deal with clusters of different distributions, and for retrieval, it allows us to improve accuracy and reduce hubs and orphans. For example, the solid line on Figure 1 depicts the counts after applying 50-Normalization. In this case, $r_o = 0.0005$ and $r_h = 0.0095$. Orphans are almost discarded by the N-normalization which compensates for the fact that the neighbors of the orphans are by definition loosely distributed. Hubs are also reduced because the distance between a hub and a given entry has to be small with respect to the distances to their respective N-Neighbors to stay small after normalization.

In [3], N is set a priori for convenience to a small value ($N = 7$). In the experiments reported in [4], the authors observed that, after a given value ($N = 25$), increasing N did not improve nor decreased significantly the accuracy. This value was then chosen by the authors for all reported evaluations.

Even though such arbitrary setting may be convenient, it is in fact counter intuitive as far as theory is concerned. As stated in the introduction, one usually wants to ensure that: *"one item of a given class has its closest neighbors belonging to the same class"*. A quantitative reformulation of this statement is to maximize the inter-class distance and minimize the intra- class distance. In this case, N should be chosen so that i_N is most of the time at the boundary of the class which includes element i.

To illustrate this, let us consider a synthetic dataset of 30 classes each of 40 2-dimensional points whose centroids are equally distributed over a diagonal, *i.e.* the coordinates of the centroids are $(1,1),(2,2),(3,3),....$ Within each class, the points are distributed around their centroids following a Gaussian distribution of standard deviation equal to 0.25. Figure 2(a) depicts with solid line the accuracy as a function of N after applying N-normalization. In this case, setting N as a low value is harmful as far as accuracy is concerned, and the maximal performance is reached when N is around the number of elements within each class.

When dealing with realistic data, several phenomena can influence the optimal N setting. The presence of outliers supports considering a smaller N than the number of elements within each class. Let us consider a sampling of the real dataset described in Section 6 composed of 11 classes each of 55 elements. The solid line on 2(b) depicts the accuracy which reaches a maximum at $N = 30$ which is a lower than the number of elements per class.

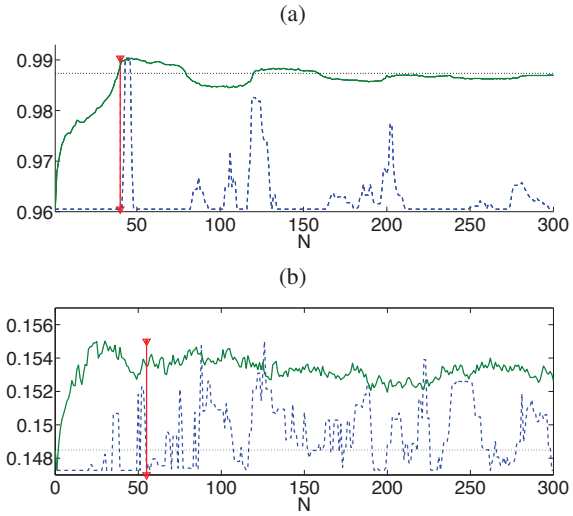

Fig. 2. *Accuracy with (solid line), without N-normalization (dotted line) and inconsistency indicator value (dashed line) as a function of N over an artificial dataset (a), a real balanced dataset (b). The indicator curve is unrelated to Y-axis values and have been rescaled for readability. The average number of elements per class is plotted as a vertical line.*

5. DETERMINING N

As shown in the previous section, the optimal N is function of the geometry of the dataset. Even in noisy and unbalanced settings, a value a bit lower than the mean cardinality of the classes seems to be a good choice. However, in practical settings, this piece of information is unavailable. One then needs to estimate N according to some relevance criterion computed solely over the data at hand.

Intuitively speaking, in a well organized dataset, nearest neighbors will "see the world" in a consistent manner. That is, if 2 elements i and j are close, their distances to any other element k should be about the same. Hubs are elements that are arbitrarily close to a large number of elements. So, in the case of hubs, this assertion does not hold anymore as this would imply that every element would be a hub. The same reasoning applies to orphans.

5.1. Inconsistency criterion

We propose the following criterion for quantifying how well organized the studied dataset is given a distance function d :

$$I_d(N) = \sum_{k=1}^{S} \sum_{j=1}^{S} \left(\frac{d_N(k,j) - d_N(k_m,j)}{d_N(k,j) + d_N(k_m,j)} \right)^2 \quad (3)$$

where k_m is the closest neighbor of k. In well organized datasets, $I_d(N)$ is low and will increase in the presence of hubs and orphans. On Figure 3, a local minima can be observed for $N = 40$, which corresponds to the number of elements within each class. However, in realistic settings it is not trivial to automatically detect such local minima.

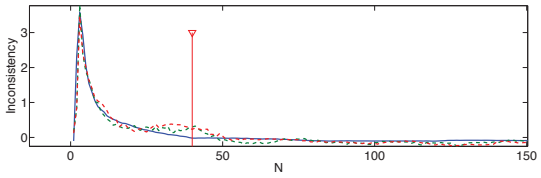

Fig. 3. *Inconsistency criterion versus N for the artificial dataset (solid line) and 2 corresponding null distributions (dashed lines). Curves have been centered for readability purposes.*

5.2. Testing against null distributions

As proposed in [5] for determining the number of clusters, it is more robust to standardize the criterion (I in our case) by comparing it with its expectation under an appropriate null reference distribution of the data.

In order to generate such null reference in the feature space, typically a Monte-carlo sampling is performed. Though, in our setting, the dimensionality of the feature space is unknown. We therefore propose to generate the null distance matrices by randomly permuting the distances.

$$d^b(i,j) = d(r_b(i), r_b(j)) \quad (4)$$

where r_b is a randomly generated permutation vector.

This allows us to have null distance matrices which have the same distribution and therefore the same intrinsic dimension without any structural information left. In such setting, it is therefore intended that the N-Normalization will not have positive effects at specific values of N. This is illustrated on Figure 3 by the 2 dotted curves which show the inconsistency for 2 null distance matrices. Their minimal value have been set to 0 for readability purposes. For those 2 curves, no significant minima can be observed. We therefore consider the following normalized inconsistency criterion:

$$NI_d(N) = 1/B \sum_{b=1}^{B} log(I_{d^b}(N) - log(I_d(N)) \quad (5)$$

where B is the number of null distance matrices, set to 2 in the experiments reported in the paper. In order to reduce spurious maxima, an order 10 median filtering is applied to $NI_d(N)$ as a post-processing step. For illustration purposes, NI_d is depicted with a dashed line on Figure 2. In the synthetic case, there is an almost perfect correlation between the optimal N, the number of elements per class and the first and maximal peak of $NI_d(N)$. In more realistic settings, the correlation with the number of elements per class is lost. However, there is still a good correlation between high values of $NI_d(N)$ and high accuracy. We therefore propose $argmax NI_d(N)$ as an estimate for the optimal N.

6. EXPERIMENTS

Unless stated otherwise, the publicly-available Magnatatune dataset is considered[1]. It is composed of 5393 songs. Each of those songs are split into 30-second audio chunks that have been tagged with a large vocabulary by the community. In order to assign a tag of genre to each song, we proceed as follows. First, a smaller vocabulary is

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

Balanced	QBE	25-Norm	A-Norm	Opt-Norm
5-AGP	0.274	0.284	0.286	0.287
n_h	114	47	54	50
r_o	0.027	0.0005	0.0009	0.0014
Unbalanced	QBE	25-Norm	A-Norm	Opt-Norm
5-AGP	0.363	0.359	0.366	0.366
n_h	113	46	55	53
r_o	0.0259	0.0005	0.0017	0.0015
Mirex	QBE	25-Norm	A-Norm	Opt-Norm
5-AGP	0.465	0.479	0.481	n-c
AFS	45.84	46.54	46.6	n-c
ABS	0.94	0.97	0.968	n-c

Table 1. *Results for balanced and unbalanced sampled datasets taken from the Magnatagatune dataset and Mirex 2010.*

extracted, containing only the tags that are explicitly referring to a musical genre. For each song, we build the list of the genre tags assigned to each of its audio chunks. The genre tag for each song is then assigned by majority voting. The resulting dataset is very unbalanced, since the mean and standard deviation of the number of elements per class are respectively about 360 and 434.

In order to evaluate the approach proposed in this paper, we consider an open-source implementation for the reference QBE. It is built using the Marsyas framework[2] that implements a feature set that has shown state-of-the-art performance in the various classification and retrieval tasks in the last MIREX[3] The distance between 2 songs is then defined as the euclidean distance between their respective normalized feature vectors.

In order to gain statistical relevance, the dataset is sampled into balanced and unbalanced smaller partitions of 2000 elements. To create a balanced partition, we seek for the largest set of classes that have their cardinality equal or superior than S divided by the number of those classes and randomly select elements within those. To create an unbalanced dataset, some elements are randomly picked from the original dataset, roughly keeping the same distribution of elements within each class as the original dataset. 100 sampled dataset are generated and used to compare the different approaches. $I_d(N)$ is computed for N up to 200[4]. 25-Norm is used for reference, A-Norm is the Adaptive normalization that considers an N_A that maximizes $I_d(N)$. Opt-Norm is the N-Normalization with N_{opt} maximizing the 5-AGP. The latter can therefore be considered as an upper bound that can only be computed when class labels are available.

As can be seen on Table 1, A-Norm improves upon the 25-Norm as far as accuracy is concerned, both for balanced and unbalanced datasets. However, 25-Norm reduces better unwanted phenomena, even more than Opt-Norm, meaning that minimizing those phenomena does not necessarily improve the retrieval performance. This might be due to the fact that the metrics considered, such as n_h, do not consider if the hub is in fact a bad hub, *i.e.* an element close to elements of many classes or a good hub, *i.e.* an element close to many elements of its class.

The proposed approach has been submitted to MIREX 2010 in order to evaluate it on an unknown dataset and to determine if the N-Normalization is relevant from an end user perspective. As can be seen on the bottom of Table 1, the N-normalization is relevant for enhancing the AGP objective measure and more importantly the AFS and ABS subjective measures. Furthermore, optimizing the value of N using the proposed method is beneficial as far as the AGP and AFS are concerned.

[2] `http://marsyas.info`

[3] Spectral Centroid, Rolloff, Flux and the Mel- Frequency Cepstral Coefficients (MFCC) as well as features related to rhythm and pitch.

[4] Other sampling strategies based on prior knowledge or heuristics can be considered in order to reduce the computational complexity.

7. CONCLUSION

In this paper, we investigated the use of the N-normalization for improving the similarity between musical objects. More specifically, a method was proposed to determine N by considering a new inconsistency criterion computed solely over the data at hand without knowledge of the geometry of the dataset at hand. From synthetic datasets to realistic datasets with balanced and unbalanced geometries, the proposed approach is useful for improving retrieval both from an objective and subjective perspective. Future work will include a more in depth study of the undesired properties that are hubs and orphans. In particular, defining new objective measure that is able to discriminate amongst good and bad hubs and orphans.

8. REFERENCES

[1] J.-J. 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.

[2] M. Radovanovic, A. Nanopoulos, and I. Mirjana, "Nearest Neighbors in High-Dimensional Data : The Emergence and Influence of Hubs," in *Proc. of the 26th International Conference on Machine Learning*, 2009.

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

[4] T. Pohle, P. Knees, M. Schedl, and G. Widmer, "Automatically Adapting the Structure of Audio Similarity Spaces," in *Proc. of the 1st Workshop on Learning the Semantics of Audio Signals*, 2006, pp. 66–75.

[5] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society*, vol. 63, no. 2, pp. 411–423, 2001.

[6] R. Foucard, J.-L. Durrieu, M. Lagrange, and G. Richard, "Multimodal Similarity between Musical Streams for Cover Version Detection," in *Proc. of ICASSP*, 2010.

[7] J. Serra, M. Zanin, C. Laurier, and M. Sordo, "Unsupervised Detection of Cover Song Sets: Accuracy Improvement and Original Identification," in *Proc. of the 10th ISMIR Conference*, 2009, pp. 225–230.

[8] M. Lagrange and J. Serra, "Unsupervised Accuracy improvment for Cover Song Detection using Spectral Connectivity Network," in *Proc. of the 11th ISMIR Conference*, 2010, pp. 595–600.

[9] W. Jin, A. K. H. Tung, J. Han, and W. Wang, "Ranking Outliers Using Symmetric Neighborhood Relationship," *Proc. of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 2006.