LEARNING AND CLEAN-UP IN A LARGE SCALE MUSIC DATABASE

ABSTRACT

We have collected a database of musical features from radio broadcasts and CD collections ($N > 10^5$). The database poses a number of hard modelling challenges including: Segmentation problems and missing and wrong meta-data. We describe our efforts towards cleaning the data using probability density estimation. We train conditional densities for checking the relation between meta-data and music features, and un-conditional densities for spotting unlikely music features. We show that the rejected samples indeed represent various types of problems in the music data. The models may in some cases assist reconstruction of meta-data.

1. INTRODUCTION

Access to large music databases including rich musical features and fat meta data is essential for research in music information retrieval, see the proceedings of the International Society for Music Informatics Retrieval conferences (ISMIR) [1] for details and background. For financial and copyright reasons there are relatively few such data bases around and they are quite limited in size. We have developed a strategy that will produce and maintain a large database for public distribution based on radio station recordings using the 'StationRipper' software [2]. We respect the copyright issue by capturing a rich set of features that have proved useful for music information retrieval - but does not allow reconstruction of a useful music signal. StationRipper produces MP3 and basic meta data (an estimate of artist and title). We also use external meta databases such as, e.g., MusicBrainz [3] to clean the acquired meta-data. We have obtained in excess of 10^5 songs with this design. The data acquired has shown relatively high quality, however, substantial amounts of cleaning is necessary due to ripping errors, data transfer issues, and stream segmentation problems. The paper is organized as follows: First we discuss the acquisition and basic cleaning steps we use for inclusion in the database. Next we describe our modelling framework incorporating both supervised and unsupervised learning steps to handle genre classification and outlier detection respectively. We discuss outlier detection in both conditional distributions (i.e., with assumed known genre context) and in unconditional distributions (with missing genre context). Our results are promising and we are currently planning the distribution of cleaned data sets.

2. DATA ACQUISITION, BASIC CLEANING AND REPRESENTATION

It is a central aim of the Danish 'intelligent sound' project[1] to create interactive demonstrations and furthermore, we are committed to establish research databases for audio modelling, in particular for music information retrieval. We consider three major sources for music data: Syndication of personal music collections, free download sites for music, and web radio stations. Here we will report on issues related to the integration and cleaning of this database.

Stationripper [2] is an application for listening, radio station navigation, and recording of music broadcasted over the internet. StationRipper stores music as MP3 files. It is programmed by Greg Ratajik and John Clegg and was first released in December 2003. At present, the latest version is 2.50 from September 2006. This paper is based on version 2.33, build March 1, 2006. The registered so-called 'Gold-version' is able to rip simultaneously up to six hundred stations, a number, however, which is crippling wrt. CPU power and bandwidth for most systems. In the experiment we report on here we have ripped up to 70 radio stations simultaneously. We have selected radio stations so as to reduce DJ voice-over, commercials, noisy ID3 tags, and other systematic errors. Web radio stations transmit at widely different bit rates. We have put a lower limit at 64 kbit/sec. The role of the bit rate for the (MFCC) feature quality has recently been discussed in [4]. Most of the songs recorded are in fact received in 128 kbit/sec or more. Thus, in principle it is possible obtain large amounts of music data with rather simple means to. Bandwidth, storage and stability however limits our effective rate. The productivity during the reported experimental campaign peaked at around 8000 songs pr. day.

We transfer songs to the music information retrieval (MIR) database according to a set of basic inclusion criteria: First, the song must be longer than 20 sec and no longer than 1200 sec. The upper limit is necessary to eliminate occasional segmentation problems with the StationRipper software. Secondly, the song should have information in the ID3 tag; at least we require song title and artist. Finally at this level, the song must not already be in the database. In our current model this implies that the artist, title and length is checked (length with 20 msec precision). Different versions, remixes etc. that appear with different lengths are included. These criteria imply that the actual number of songs included is further reduced from the raw numbers mentioned above.

After checking with the basic criteria the song is processed and uploaded to a database using our newly developed

[1] www.intelligentsound.org

Winamp plug-in [5]. The plug-in computes three representations. The basic representation is the set of mel frequency cepstral coefficients (MFCCs) based on a 20 msec analysis window with 50% overlap. For details on MFCC estimation, see e.g., [4]. This representation creates about 0.5Mb pr. 60sec of a song. The second level representation is based on temporal integration using the multi variate autoregressive (MAR) approach of [6]. This 135 dimensional feature vector is estimated in 1sec windows. We check that the length of the MAR's actually match the length of the song (accept range: 94-105%).

Finally, as the third level in the processing pipeline we perform supervised kernel-projection of the MAR vectors to form a relatively low-dimensional ($D = 15$) feature vector that implements a basic musical genre indicator (GI) [7]. The GI's are trained on a 12.000 clip dataset with high-quality meta-data, in which each dimension corresponds to a genre from the `Amazon.com` genre set. This genre definition is different from the set used in our database in general. The complete set of MFCC's, MAR's and the 15-dimensional *song-average* GI's are stored in the database. This rather rich 'fingerprint' has been cleared with the national copyright owners organization (KODA). It is not possible to reconstruct a useful representation for listening from the fingerprint.

At the time of writing the total number of unique songs in the database is $N = 103644$, the unique artists and titles amounts to $N_u = 92235$. The StationRipper part consists of $N_s = 62100$ songs.

We here focus on modelling in the 15-dimensional GI representation. To understand better the nature of the database we consider sets that have two different origins: A dataset which is obtained from private collection syndication *with* genre labels, and the StationRipper database which has only artist and title labels.

2.1 Additional metadata recovery using MusicBrainz

StationRipper typically provides rather shallow metadata consisting of song title and artist name, hence, lacking important information such as year, genre, album, etc. Furthermore, the data set is somewhat biased towards certain genres. Although there exists radio channels focussed on jazz, classical and folk, the main body of stations are labelled rock, pop, dance or various forms of electronica.

MusicBrainz is a comprehensive public community music meta-database. The MusicBrainz data can be accessed either through the web site, or with client programs. MusicBrainz can be used to identify CDs and provide information about the CD, about the artist or about related information. We primarily aim to use MusicBrainz to clean up meta-data tags. In a preliminary screening we found that about 35% of our database songs had exact (artist+title) matches in the MusicBrainz database. The potentially useful tag 'year' was present in a subset of size 25% of the songs.

3. OUTLIER MODELING

We use a combination of supervised and unsupervised learning to model and clean the database. The distributional properties of the reduced dimensional data is illustrated in figure 2. The relatively complex distributions motivate the use of flexible density models in the GI space.

3.1 Parzen window estimators

Parzen window density estimators are well suited for outlier detection because they typically create compact pdf's. Densities can alternatively be approximated by mixture models, see e.g., [8], however they tend to produce wider, hence, less specific distributions, which may be desirable for other tasks such as meta-data generalization.

In [9] we proposed to detect outliers in a meta-data conditional sense by estimating class *conditional* probability density functions (pdf's). Here a relevant meta-data is the genre label. The class conditional densities can be used to locate items with novelty in the data/meta-data relation. The less specific *un-conditional* pdf's can be used to spot novelty in data for which we have no meta-data.

The Parzen window model is based on a training set of data $D = \{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_N\}$ of size N

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} p_0(\mathbf{x} - \mathbf{x}_n|\sigma^2) \quad (1)$$

Where $p_0(.|\sigma^2)$ is a simple normalized isotropic Gaussian kernel with variance parameter σ^2. The variance parameter is estimated by a leave-one-out procedure, see e.g., [9] for details. Genre conditional densities $p(\mathbf{x}|\text{genre})$ are estimated from training sets solely from a the given genre label. In the two data sets we consider, we only have labels in one set and these labels come with the data in the syndication process. Therefore it is a separate important issue to check the label consistency.

4. RESULTS

We will illustrate outlier detection and conditional outlier detection in experiments on subsets of the database consisting of songs acquired using the labelled 'syndicated data' and the un-labelled StationRipper software. Here we will first train conditional density models on the labelled data and clean for inconsistencies between labels and music features. Next, we will estimate the un-conditional density and clean the unlabelled data.

4.1 Experimental design

As training data for the analysis we extracted four representative genre sets (*rock, jazz, dance, classical*) from the part of the syndicated database in which the features originate from personal music collections. For this subset we are confident that the music files are actually music, i.e., has few problems with segmentation, commercials etc. On the other hand the genre labels can be unreliable. We select training data from these subsets.

4.2 Label to music consistency

The first experiment concerns the conditional density $p(\mathbf{x}|\text{rock})$ based on a rock sub-sample $N_r = 1000$. We trained the kernel estimator width using a leave-one-out Newton scheme to obtain $\sigma^2_{\text{opt}} = 0.17$. For comparison, the mean square distance between members in the training set is 0.19. In the top panel figure 1. we show a histogram of the pdf-values obtained for the Part 1. rock test set ($N_{rt} = 5400$).

To illustrate the specificity of the density estimators we evaluate the pdf-values for a subset of data labelled 'classical'. If this test set was indeed all proper classical music

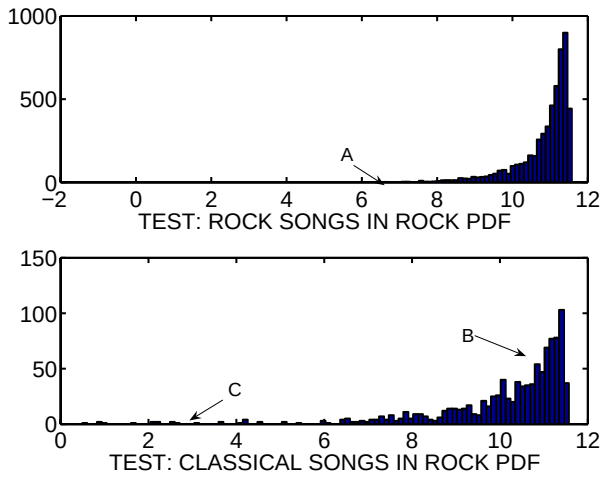

Figure 1: Top panel: Histogram of pdf-values of rock test songs in the conditional density $p(\mathbf{x}|\text{rock})$. Bottom panel: Histogram of pdf-values of test songs labelled classical measured with the conditional density $p(\mathbf{x}|\text{rock})$. (A) are unlikely rock songs, (B) are classical songs that are likely in the rock context, while (C) are songs that are novelty relative to the rock context.

	Author	Title
1	Dirty Dancing	Bill Medley and Jennifer Warn
2	Jethro Tull	Quizz Kid
3	U2	Staring at the sun
4	Genesis	One Man's fool

Table 1: List of songs from the set labelled classical that are *likely* under the rock pdf. This novelty set consists of rock songs misclassified as classical, here we could not only spot errors but also potentially re-label the songs appropriately as rock.

it should be rejected under the rock pdf. The histogram of this smaller set ($N_{ct} = 1014$) is shown in the bottom panel of figure 1. Surprisingly, we see that a large portion of the densities are of similar magnitude as typical pdf-values in the rock test set. To understand this we inspected the list of high rock-pdf value entries, which according to the label should be classical music. This list, which is shown in table 1, reveals that there is a massive misclassification problem with the 'classical' labels of this set. Indeed many of these songs would be more correctly labelled rock.

We also checked songs from the classical labelled set that have very low probability under the rock-pdf. The list in table 2 shows that these entries are indeed classical music, hence, should properly be rejected under the rock-pdf.

In table 3. we list a few of the entries that are novelties in the rock-pdf, but have genre label rock. The list contains non-rock pop and rap, however, one song by 'Bottle Rockets' is also listed. This may be a false alarm, or a yet unknown error type.

4.3 Cleaning for outliers in the overall density

In this experiment we simulate cleaning a data set without labels. We combine the models trained above using even a priori class probabilities to get the joint pdf. In table 4. we list outlier examples from the larger unlabelled StationRipper dataset rejected under this joint pdf. Here the list is topped by music downloaded from radio stations that have various technical issues. Music from these stations has subsequently been deleted from the database. In figure 2 these outliers are seen as a cluster of points in the right panel of unlabelled data located away from the main 'axis aligned' groups seen in the left panel.

	Author	Title
1	Chopin	Waltz 6 (Minute Waltz)
2	Several Orchestras	Jesus bleibet meine Freunde
3	Chopin	Waltzer (Tempo Giusto)
4	Rachmaninoff	Barcarolle in G

Table 2: List of songs from the set labelled classical that were unlikely under the rock pdf. The set contains songs that are correctly labelled classical, hence should be accepted.

	Author	Title
1	A Camp	Angel of Sadness
2	Bottle Rockets	Radar Gun
3	Zindy Kuku Boogaloo	Mr. Big Stuff
4	Thomas Helmig	Lovers And Friends

Table 3: List of novelty songs in the conditional density estimator for rock music. The list of songs that are novelty under the rock pdf is topped by pop songs and a rap-hip hop entry. The 'Bottle Rockets' song may be a 'false alarm'.

5. CONCLUSION AND DISCUSSION

We have outlined steps towards acquisition, learning and cleaning of a large scale public MIR database.

We have outlined an approach to outlier detection using non-parametric density estimators. We have previously used a similar approach in a neuroinformatics application [9]. In this application we also investigated parametric Gaussian mixture densities, however, they tend to provide too dispersed probability density functions.

Alternatives to this approach are found in the datamining literature, in [10, 11], a heuristic is proposed based on a distance measure, basically enough neighbors need to be in a certain distance. The present approach shares many aspects with this method, however, using an optimized density model we make sure that the decisions are statistically well founded.

Recently an approach based on density estimation as we advocate here and earlier in [9], was developed in the VLDB contribution [12]. However, using a different kernel and a heuristic for estimating the width based a scaling rule involving the coordinate wise standard deviations. In many data sets the coordinate standard deviations do not well summarize the underlying distribution, c.f., figure 2.

In this work we have investigated density based outlier detection for both labelled and unlabelled data. For conditional densities appropriate for labelled data we can test the consistency of the label and music feature vector. For the unconditional density we can test whether a given music feature vector is likely to represent music similar to that of the training database.

In our StationRipper based collection scheme we are typ-

Rank	Author	Title	Radio	comment
1	81702 Delerium	Just A Dream	Radio Paradise	A
2	Talking Heads	Houses in Motion	Radio Paradise	A
3	Toad The Wet Sprocket	Something‘s Always Wrong	Radio Paradise	A
4	Not Complete DJs	Destination [Original Vocal Mix]	radioparty.pl	A
5	Hi	Per - Gimme More (Club Mix)	radioparty.pl	A
6	Billy Joel	You’re Only Human (Second Wind)	Atlantic Sound Factory	A
7	Sahin Gultekin	Kalenin Bedenleri	www.radyoiz.com	A
8	Morcheeba	World Looking In	Radio Paradise	A
9	The Dears	Who Are You, Defenders Of The	Radio Paradise	A
10	Badly Drawn Boy	The Shining	Radio Paradise	A
75	George Hamilton IV	A Rose and a Baby Ruth	MyMixRadio	B
100	Fleetwood Mac	Tusk	Atlantic Sound Factory	A
150	Sarah Vaughan	C’Est la Vie	MyMixRadio	B
350	Youngbloodz	Damn G	1.FM Jamz	OK
351	DJ Dr. Dubbs	Battle of the Beats	1.FM Jamz	DJ
1000	D	Tek vs Cyrus The Virus - Dare	Digitally imported Goa	OK

Table 4: List of outliers/novelty songs under the unconditional density estimator. The list of songs that are novelty under the genre global pdf is topped by songs that have a technical issue (A) that lead the plugin to produce invalid MFCC data. The problem we so have spotted has led to deletion of data from several radio stations, including the station ‘Radio Paradise’. Further down the list we find songs that have no technical issues, but have are in genres not considered in the present context (*crooners*, B). In position $P = 1000$ we find a dance song for which we have found no issues, hence, possibly representing a false alarm.

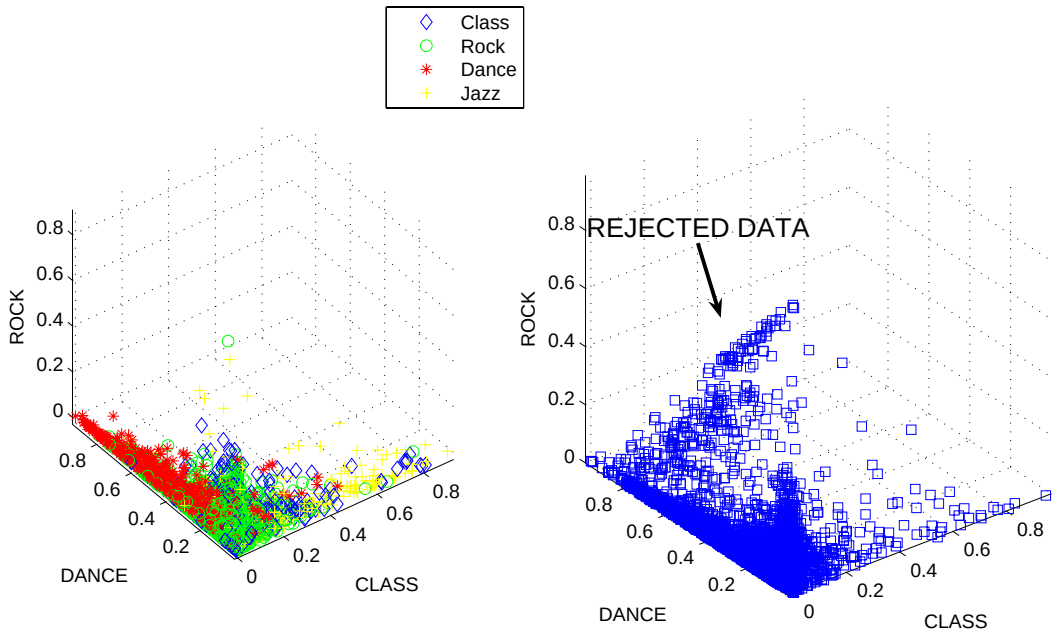

Figure 2: Left panel: Scatter plot of the CD-collection data subset ($N_{CD} = 9100$) in the genre indicator (GI) dimensions *Classical*, *Dance*, *Rock*. Right panel: Scatter plot of the StationRipper data ($N_s = 62100$) in the genre indicator (GI) dimensions. The cluster of points indicated by the arrow ‘REJECTED DATA’ in the right panel are rejected as outliers in the density model estimated from the data in the left panel.

ically provided with a music feature vector, an artist name and a song title. A natural sequence would be to first test whether a given music feature vector passes the unconditional pdf test, i.e., is music. If we check with a meta-database, say MusicBrainz, and find a label then we can test whether is sufficiently probable in the given genre. If the song fails the latter, it may be re-labelled using the model density, discarded or subjected to a manual listening test. As some meta-databases base the genre label on artist identity we expect quite a number of mislabelled songs from artist in cross over genres.

In our experiments data that was syndicated from private music collections turned out to have genre label errors, which was spotted by use of genre *conditional* density estimation. Ripping web radio is a route to very large data sets. However, these data do not immediately provide meta-data. By testing with global density estimation, we found radio stations that produced technical problems for a data collection pipeline, the source of these remain an issue for our programming team at present. Sofar these recordings have been deleted from the database.

In general the density model approach shows promise for cleaning and may also be used for bootstrapping genre labels from small carefully labelled sets.

REFERENCES

[1] www.ismir.net.

[2] www.stationripper.com.

[3] www.musicbrainz.org.

[4] S. Sigurdsson, K. B. Petersen, and T. Lehn-Schiøler, "Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music," in *Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR)*, 2006. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?4690

[5] T. Lehn-Schiøler, J. Arenas-Garca, K. B. Petersen, and L. K. Hansen, "A genre classification plug-in for data collection," in *ISMIR*, 2006. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?4520

[6] A. Meng, "Temporal feature integration for music organisation," Ph.D. dissertation, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, 2006, supervised by Jan Larsen and Lars Kai Hansen, IMM. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?4502

[7] J. Arenas-Garca, K. B. Petersen, and L. K. Hansen, "Sparse kernel orthonormalized pls for feature extraction in large data sets," in *Advances in Neural Information Processing Systems 2006 (Proc. to appear)*, 2006.

[8] A. Berenzweig, D. Ellis, and S. Lawrence, "Anchor space for classification and similarity measurement of music," in *Proceedings of IEEE International Conerence on Multimedia and Expo (ICME)*, 2003.

[9] F. Å. Nielsen and L. K. Hansen, "Modeling of activation data in the brainmaptm database: Detection of outliers," *Human Brain Mapping*, vol. 15, no. 3, pp. 146–156, mar 2002. [Online]. Available: http://www3.interscience.wiley.com/cgi-bin/abstract/89013001/

[10] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *SIGMOD Conference*, 2001. [Online]. Available: citeseer.ist.psu.edu/aggarwal01outlier.html

[11] F. Angiulli and C. Pizzuti, "Outlier mining in large high-dimensional data sets," *IEEE Trans. Knowl. Data Eng.*, vol. 17, pp. 203–215, 2005.

[12] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, "Online outlier detection in sensor data using non-parametric models," in *VLDB*, 2006, conf, pp. 187–198.