IMPROVING AUTO-TAGGING BY MODELING SEMANTIC CO-OCCURRENCES

Riccardo Miotto
University of Padova
miottori@dei.unipd.it

Luke Barrington
UC San Diego
lukeinusa@gmail.com

Gert Lanckriet
UC San Diego
gert@ece.ucsd.edu

ABSTRACT

Automatic taggers describe music in terms of a multinomial distribution over relevant semantic concepts. This paper presents a framework for improving automatic tagging of music content by modeling contextual relationships between these semantic concepts. The framework extends existing auto-tagging methods by adding a Dirichlet mixture to model the contextual co-occurrences between semantic multinomials. Experimental results show that adding context improves automatic annotation and retrieval of music and demonstrate that the Dirichlet mixture is an appropriate model for capturing co-occurrences between semantics.

1. INTRODUCTION

A central goal of music information retrieval (MIR) is to create systems that can efficiently and effectively retrieve songs from massive music collections. A potential solution to this challenge is to describe songs with a collection of manually annotated meaningful words (tags) and to perform retrieval based on these text descriptions. Commercial recommendation systems such Last.fm [1] and Pandora [2] extensively use this semantic similarity approach to create recommendation lists. Tags are useful because they contextualize a song by describing human emotions, personal style, geographic origins, spiritual foundations, historical period, or particular uses of the song.

1.1 Auto-Tagging

The continuous growth of music collections is making manual human annotation of every song infeasible. In response, several scalable approaches have been proposed for labeling music with semantics including social tagging [7], web mining [6] or tag propagation from similar songs [12], each with advantages and disadvantages [14]. In particular, MIR researchers have proposed content-based "auto-taggers" – methods that analyze acoustic waveforms and automatically assign meaningful words to songs. Much of this work has been inspired by related methods for automatic image annotation [13].

One of the the first proposed approaches used Gaussian Mixture Models (GMM) computed over the audio features of the training examples to represent a vocabulary of words [15]. An alternative model, the Codeword Bernoulli Average (CBA) [5] attempted to predict the probability that a tag applies to a song based on a vector quantized representation of the audio signal. Regardless of the model used, the output of an auto-tagger is a vector of tag probabilities which may be interpreted as a *semantic multinomial* (SMN), a distribution that characterizes relevance of each tag to a song. Semantic multinomials capture patterns in a song's waveform that represent high-level properties such as genres, emotions or instrumentation.

1.2 Tag co-occurrence

Auto-tagging models aim to capture statistically regular patterns in the audio content and associate these patterns with descriptive semantics. In general, these models treat each tag *independently*, ignoring the *context* that derives from associations between tags. Indeed, while some semantic associations in music are inspired by direct auditory cues (e.g., hearing a "violin"), others are inferred through contextual relationships (e.g., inferring "cello" and "bassoon", when listening to "orchestral classic music"). This gives rise to statistically significant co-occurrence patterns of semantic concepts in the training data (e.g., many "rock" songs also tagged as "loud"), and thus in the SMNs. We suggest that actively capturing correlations in SMNs can improve the semantic description of a song.

Two situations cause tags to co-occur in semantic multinomial distributions. The first is when a tag accidentally co-occurs with another concept. Accidental co-occurrences could be due to many reasons, ranging from poor posterior probability estimates arising from auto-tagger errors, to the unavoidable ambiguous interpretation of music, such as confusing "trumpet" and "trombone". The second type of tag co-occurrence results from feature vectors that truly describe multiple musical concepts. For example, a "cello" piece is very likely to have feature vectors that also fit tags such as "classical music" or "violin". While only co-occurrences of the second type are indicative of *true* contextual relationships, SMN distributions derived from acoustic content exhibit both types of co-occurrences.

[1] http://www.last.fm
[2] http://www.pandora.com

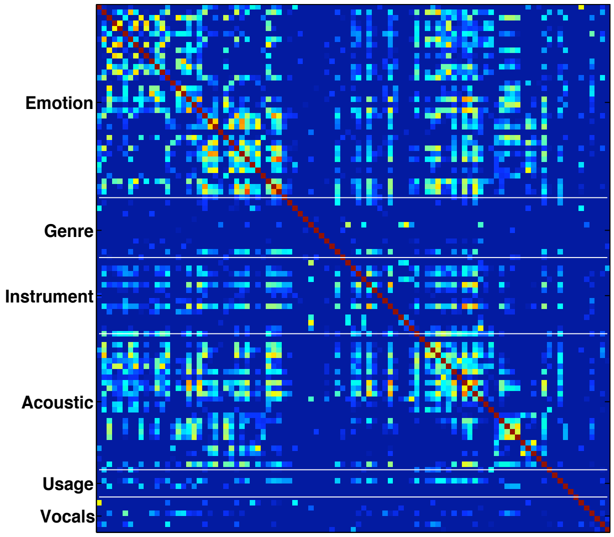

Figure 1. Co-occurrence patterns for *CAL500*; redder points imply high correlation between tags.

To understand the extent of tag co-occurrences, we examine the Computer Audition Lab 500 (CAL500) dataset, used later in our experiments (see Section 4 for more details). Figure 1 depicts the pairwise correlation matrix between CAL500 tags. Correlation values have been computed through an application of Jaccard's Coefficients [8],

$$n_{ij} = \frac{P(w_i \cap w_j)}{P(w_i) + P(w_j) - P(w_i \cap w_j)}\ , \qquad (1)$$

which provide a measure of the strength of the association between the general words w_i and w_j, normalized by the total number of times the two words appear. The n_{ij} coefficients range between 0 and 1, with $n_{ij} > 0$ if the tags are not mutually exclusive (i.e., if they occur together in some songs). In Figure 1, redder parts represent tag pairs that are highly correlated (i.e., where n_{ij} is large). As can be seen, correlation is present in many tags and it is particularly prevalent in the "Emotion" and "Acoustic" categories whereas tags categorized as "Genre" display few correlation patterns.

The co-occurrence patterns illustrated in Figure 1 are not explicitly captured by auto-taggers that model acoustics independently for each tag. Although SMNs capture patterns at the song level that are predictive of semantic tags, each dimension of the semantic space (i.e., each tag) is assumed to be independent from all others. Exploiting these regular co-occurrences - giving the semantics context - could provide a better semantic description of music.

This suggests an extension of auto-tagging models by adding one additional layer of semantic representation that explicitly captures tag co-occurrences. We began by modeling the probability distribution of tags given audio features, placing each song in a semantic space. Now, by modeling a probability distribution of the SMNs derived from each song - a distribution over distributions - we can obtain a richer semantic description. We refer to these representations as *contextual models*.

1.3 Modeling Context

In this paper, we present a novel approach to automatically tagging music with descriptive words by thinking of each semantic concept as defining a broader *context* that causes multiple, related tags to co-occur in the description of a song. For each tag, we learn a Dirichlet mixture (DM) to model the distribution of the SMNs derived from all training songs for that tag. This DM-based "contextual tag model" is inspired by similar work on modeling the semantics of images [11] where it was proposed as a framework for combining object-centric and scene-centric methods to model contextual relationships between visual concepts. The DM can robustly infer contextually meaningful co-occurrence patterns between tags in semantic multinomials, while removing accidental co-occurrences that might be present in some of the individual song-level SMNs.

2. RELATED WORK

Some recent work in music information retrieval has exploited tag correlation and context. Yang et al. [16] formulate tag detection as an ordinal regression problem to explicitly take advantage of the ordinal relationship between concepts. Moreover, they proposed to leverage the co-occurrence patterns of tags for context fusion and employ tag selection to remove irrelevant or noisy tags. Unlike our approach, the latter is a single-level model, incorporating the tag correlation during the training of each individual detector. Ness et al. [10] propose a hierarchy of two linear SVMs where the first classifier highlighted the audio patterns and output a vector of tag affinities (analogous to a SMN), and the second layer modeled the contextual relationships between tags. Modeling context was also proposed in [3] where a second stage used a learning and correlation reweighing scheme to boost the result of tag detection, and, earlier, in [1] where authors used a decision tree to refine the result of individual detectors.

Our approach using the DM to model context is appropriate for two reasons. First, the DM is a *generative* model that is learned from only positive training examples i.e., songs which have been positively associated with a semantic tag. Unlike discriminative models (e.g., SVMs, boosting, decision trees) which also require negative examples, generative models can accommodate weakly labeled training data where the absence of an association between a song and a tag does not guarantee that no such association exists. Second, the Dirichlet is a distribution over parameters of the multinomial distribution, making it a probabilistically appropriate model of semantic multinomials derived from auto-taggers.

3. AUTO-TAGGING WITH DIRICHLET MIXTURES

We start by briefly defining the problem and by reviewing the song-level auto-tagging system described in [15].

3.1 Problem formulation

The task of semantic annotation and retrieval can be seen as a supervised multiclass, multilabel classification problem, where each class is a word w_i from a vocabulary $\mathcal{V} = \{w_1, ..., w_{|\mathcal{V}|}\}$ of unique tags, and each song is labeled with multiple words. A song is represented as a series of audio content features, $\mathcal{X} = \{x_1, ..., x_T\}$, where x_t represents a vector of features, and T is related to the length of the audio content; the goal is to find the words $w_i \in \mathcal{V}$ which best describe a given song. Each song can then be represented as an annotation vector $\pi = (\pi_1, ...\pi_{|\mathcal{V}|})$, where $\pi_i > 0$ if w_i has a positive semantic association with the song and $\pi_i = 0$ otherwise. The coefficients π_i represent the strength of semantic association between the song and word w_i and are termed *semantic weights* [15] or *affinity values* [10].

3.2 Defining a semantic space

Various auto-tagging methods have been proposed for deriving the semantic weights from acoustic features including hierarchical Gaussian mixture models [15], support vector machines [2,10], codeword Bernoulli averaging [5] and boosting [3]. Any of these auto-taggers may be used to produce semantic multinomials — a set of semantic weights — that describe songs, a process that is illustrated on the left of Figure 2. In this work, we use the hierarchical GMM approach and briefly review it hereafter but refer the reader to [15] for the details of this model.

For each word w_i in the vocabulary, we train a tag-level probability distribution over the audio feature space, e.g. $P_{X|W}(x|w_i)$ for $i = 1, \ldots, |\mathcal{V}|$. The most relevant tags for a song $\mathcal{X}$ are the words with highest posterior probability, computed using Bayes' rule:

$$\pi_i = P_{W|X}(w_i|\mathcal{X}) = \frac{P_{X|W}(\mathcal{X}|w_i)\, P_W(w_i)}{P_X(\mathcal{X})} , \qquad (2)$$

where $P_W(w_i)$ is the prior of the i^{th} word. We assume an uniform prior, e.g., $P_W(w_i) = 1/|\mathcal{V}|$ for $i = 1, \ldots, |\mathcal{V}|$. We compute the song prior as $p(\mathcal{X}) = \sum_{i=1}^{|\mathcal{V}|} p(\mathcal{X}|w_i)p(w_i)$. We follow [15] in estimating the likelihood term in Equation 2, $P_{X|W}(\mathcal{X}|w_i)$, with the geometric average of the individual feature likelihoods of all the songs positively associated with word w_i:

$$P_{X|W}(\mathcal{X}|w_i) = \prod_{t=1}^{T} \left(P_{X|W}(x_t|w_i)\right)^{\frac{1}{T}} , \qquad (3)$$

where the distribution $P_{X|W}(x|w_i)$ is modeled as a mixture of Gaussians. The $P_{X|W}(x|w_i)$ distributions capture the patterns of audio content that are predictive of each word w_i.

Given an unseen test song, represented by a set of audio feature vectors $\mathcal{X}$, we compute the posterior probabilities for the presence of concept $w_i \in \mathcal{V}$ from Equation 2. Collecting the posterior probabilities of each word results in an annotation vector describing the song, $\pi = \{\pi_1, ..., \pi_{|\mathcal{V}|}\}$, where π_i denotes the posterior word probability $P_{W|X}(w_i|\mathcal{X})$. With appropriate normalization ($s.t. \sum_i \pi_i = 1$), this vector can be conceived of as a *semantic multinomial* (SMN) which lies on a probability simplex defined as a *semantic space*. The semantic multinomial is analogous to a *document vector* of word counts, often used in natural language processing [8], and it captures all the semantic information about the song.

3.3 A model to learn context

To capture the common patterns in the SMNs and model co-occurrences between tags, we learn *contextual tag models* in the semantic space from the SMNs of the all songs in a training set that have been labeled with each tag. This contextual modeling stage is illustrated on the right of Figure 2. Just as we modeled acoustic feature vectors as samples from a mixture of Gaussians, we consider that semantic multinomials π are drawn from a mixture of Dirichlet distributions over the semantic space [11]:

$$P_{\Pi|W}(\pi|w; \Omega^w) = \sum_k \beta_k^w \,\mathrm{Dir}(\pi|\alpha_k^w) , \qquad (4)$$

The contextual model for the word w is characterized by a vector of parameters $\Omega^w = \{\beta_k^w, \alpha_k^w\}$, where β_k is a probability mass function ($\sum_k \beta_k^w = 1$), $\mathrm{Dir}(\pi; \alpha)$ a Dirichlet distribution of parameter $\alpha = \{\alpha_1, ..., \alpha_{|V|}\}$,

$$\mathrm{Dir}(\pi|\alpha) = \frac{\Gamma(\sum_{i=1}^{|V|} \alpha_i)}{\prod_{i=1}^{|V|} \Gamma(\alpha_i)} \prod_{i=1}^{|V|} (\pi_i)^{\alpha_i - 1} , \qquad (5)$$

and $\Gamma(.)$ the Gamma function.

The parameters Ω^w are learned from the SMNs π_n of all the songs annotated with word w. Note that the contextual models $P_{\Pi|W}(\pi|w)$ play, in the semantic space, a similar role to the models $P_{X|W}(\mathcal{X}|w)$ in the acoustic feature space.

The learning process for the Dirichlet mixture model relies on the maximum likelihood estimation, via the generalized expectation-maximization (GEM) algorithm. GEM is an extension of the standard EM algorithm, applicable when the M-step of the latter is intractable. The E-step computes the expected values of the component probability distribution β_k, whereas the generalized M-step estimates the parameters α_k. Rather than solving for the parameters of maximum likelihood, each M-step simply produces an estimate of the likelihood which is higher than that available in the previous iteration. This is known to be sufficient for EM convergence [4]. Parameter estimation is achieved through an application of the Newton-Raphson algorithm [9].

Given an unseen test song described by the SMN $\pi = \{\pi_1, ..., \pi_{|\mathcal{V}|}\}$, the assignment of a word, w_i, results from a Bayes decision rule based on the posterior word probabilities in the context space:

$$P_{W|\Pi}(w_i|\pi) = \frac{P_{\Pi|W}(\pi|w_i)P_W(w_i)}{P_\Pi(\pi)} . \qquad (6)$$

Again we assume a uniform word prior probability $P_W(w_i)$. Collecting all the posterior probabilities

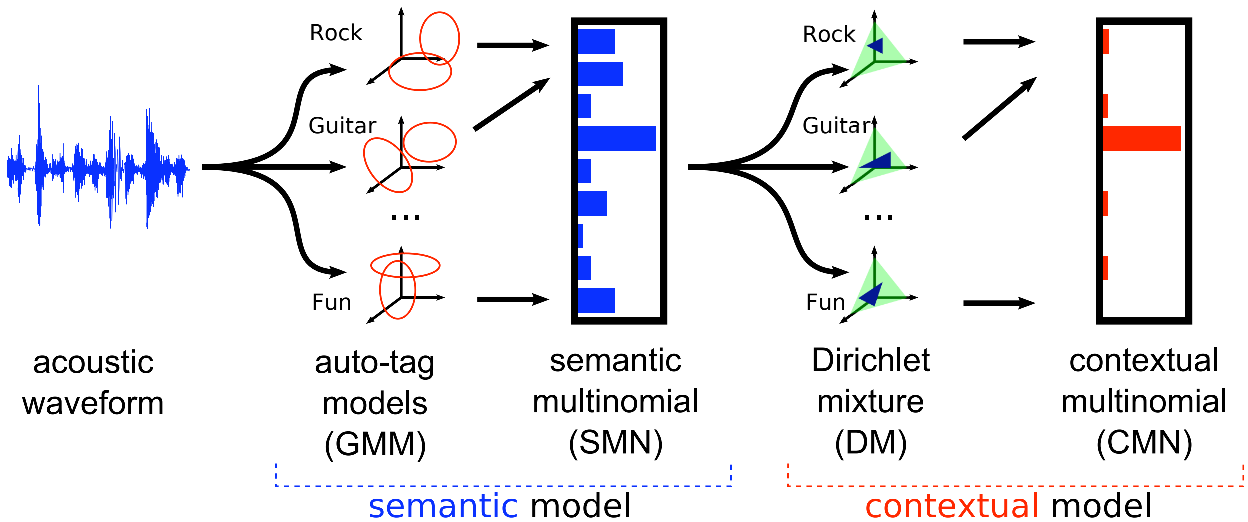

Figure 2. Overview of the system: the Dirichlet Mixture models context by considering co-occurrences patterns between auto-tags lying in a semantic space.

$P_{W|\Pi}(w_i|\pi) = \theta_i$ and normalizing ($s.t. \sum_i \theta_i = 1$), we build the vector $\theta = (\theta_1, ..., \theta_{|\mathcal{V}|})$, denoted as the *contextual multinomial* (CMN) distribution of a song. Similar to the semantic space defined in Section 3.2, CMN vectors lies in a *contextual space* (see Figure 2).

4. EXPERIMENTAL RESULTS

In this section, we demonstrate the impact of contextual models, and in particular the DM, on automatically tagging music with meaningful words.

4.1 CAL500 Dataset

The Computer Audition Lab 500 (CAL500) [15] dataset comprises 502 songs by 502 different artists. Each song has been annotated by at least 3 humans using a vocabulary composed of 174 tags from 6 different semantic categories, representing both objective and subjective concepts.

The songs are described by Mel-Frequency Cepstral Coefficient (MFCC) feature vectors; each MFCC vector summarizes the spectral content of 23ms windows of a song. Our experiments use 39-dimensional MFCC-Delta feature vectors, composed by appending the first and second instantaneous derivatives to the 13-component MFCCs.

A first analysis of the dataset demonstrates an imbalance in the distribution of tags: while frequent tags can have more than 300 positive examples, some others have less than 10 ones. This is not a big problem when training auto-taggers since each song is described by a large number of features vectors. However, the resulting set of SMNs describing songs is much smaller than the number of feature vectors and thus, we require more songs to adequately train the contextual models. For this reason, our evaluation considers only the tags with more than 30 examples, aiming to have at least $20 - 25$ examples in the training set with the remainder in the test set. This reduces the CAL500 vocabulary to 97 tags: 11 genres, 14 instruments, 25 acoustic qualities, 6 vocal characteristics, 35 emotions and 6 usages.

To provide sufficient data to train the DM, we extract multiple SMNs from each song, each derived from clips lasting 3 seconds. We find empirically that, unlike images which generally depict only a few semantic concepts (i.e., their SMNs have a few peaks that dominate all other tags), even a short music clip can be reasonably tagged with many words and the resulting SMNs tend to be much more uniform. For this reason, when learning DM models, we threshold the SMNs, retaining at most the ten largest affinity values and setting all other dimensions to zero.

4.2 Annotation and Retrieval

We evaluate auto-tagging performance on both annotation and retrieval tasks. In the *annotation* task, we use Equation 6 to label each test song with the ten most likely tags. Performance is measured using mean per-tag precision, recall and F-score. Per-tag precision is the probability that a tag used by the model is correctly applied to a song. Per-tag recall is the probability that the model annotates all the tags that should apply to a song. F-score is the harmonic mean of precision and recall, and is a single measure of overall annotation performance.

In the *retrieval* task, we rank-order all songs according to their relevance to a query tag. The retrieval goal is to have highly relevant songs at the top of the ranking list as this is the most crucial requirement in a music retrieval system. We consider the mean average precision (MAP) and the *precision at k* (k = 3, 5, 10). For completeness, we also report the area under the receiver operating characteristic curve (AROC) as a measure of the quality of the complete ranking [8].

Evaluation was performed using 5-fold cross validation, with 400 songs in the training set, and 100 in the test set. The folds were built such that each song appeared in the test set exactly once. The results reported in Table 1 display the annotation and retrieval metrics, averaged over all

		Annotation			Retrieval				
		Precision	Recall	F-Score	P3	P5	P10	MAP	AROC
Semantic	CBA	0.361	0.212	0.267	0.463	0.458	0.440	0.425	0.691
	GMM	0.405	0.202	0.269	0.456	0.455	0.441	0.433	**0.698**
Context	SVM	0.380	0.230	0.286	0.512	0.487	0.449	0.434	0.687
	DM	**0.441**	**0.232**	**0.303**	**0.519**	**0.501**	**0.470**	**0.443**	0.697
	Upper Bound	0.716	0.471	0.568	1.000	0.993	0.942	1.000	1.000
	Random	0.231	0.101	0.140	0.255	0.249	0.250	0.277	0.504

Table 1. Performance of different auto-taggers: the Codeword Bernoulli Average (CBA) and Gaussian Mixture Models (GMM) consider semantics alone whereas the Support Vector Machine (SVM) and Dirichlet Mixture (DM) models learn contextual relationships between the semantic multinomials produced by the GMM. All experiments were performed on the same songs represented by the same set of features. "Random" is a baseline that annotates and ranks songs randomly. "Upper Bound" uses the optimal labeling for each evaluation metric and shows the upper limit on what any system could achieve.

tags in the vocabulary.

4.3 Contextual improvement

The proposed contextual modeling approach is compared to some recent state of the art auto-tagging approaches: the GMM model [15] alone (i.e., without context) and the CBA model [5]. For the CBA model, each song is represented as a histogram over a codebook of 500 vector-quantized MFCCs. For each fold we trained the codebook models only on the songs in the training set. All the code was provided by the authors of [5].

We see in Table 1 that there is significant benefit from modeling context on almost all annotation and retrieval metrics. In particular, the precision-at-k metrics demonstrate improvements at the top of the ranked retrieval list but not throughout list (based on AROC). It can be argued that precision-at-k metrics consider the part of the ranked list which is most interesting for users of a semantic music retrieval engine.

4.4 DM as a model of context

The center rows of Table 1 compare the DM approach for modeling semantic co-occurrences to a Support Vector Machine (SVM). As with the DM, we trained a contextual SVM for each tag using the semantic multinomials as the input feature vector. Using SVM as a model of context was first proposed in [10] although their approach differs in the features used (median MFCC texture windows) and in the semantic model (SVM), so our results do not present a direct comparison with [10]. Our goal is simply to compare the DM and SVM as models of contextual relationships. The context SVM does not benefit from pre-processing the SMNs (results not shown), thus SVMs are trained on all the original semantic values. Table 1 shows that DM generally improves on all the metrics and never performs worse. In particular, the DM significantly improves on the SVM for the annotation precision, F-score, P5, P10 and AROC metrics (t-test, 10% significance level); all the other metrics generally improve and never perform significantly worse.

Table 2 breaks up the evaluation over the different tag categories. As can be seen, all categories but "Genre" show clear benefit from contextual modeling. Note that improvements are related to the tag co-occurrences depicted in Figure 1. In fact, all the categories showing a high degree of co-occurrences ("Emotion", "Instrument" and "Acoustic") improved with respect to the GMM. Though not exhibiting as much co-occurrence, the "Usage" and "Vocals" categories, which perform poorly using semantics alone, benefit from the de-noising effect of learning contextual relations. In these cases, the extra information from even only few co-occurrences can lead to improvements in the quality of auto-tagging. Conversely, since the "Genre" category does not exhibit much co-occurrence (i.e., genres are more exclusive), we do not gain benefit from additional contextual modeling. It has to be noted that SVM performs better for the "Genre" category, especially in the top of the ranking list; we believe that in this case SVM benefits from some de-noising effects that DM is not able to capture.

4.5 Predictive co-occurrences

Finally, we include some examples of learned contextual models for 6 tags, representing each semantic category in CAL500. Table 3 shows the top three semantic multinomial dimensions that have most influence on the contextual models for each tag. These examples illustrate how the DM uses context to improve automatic tagging by learning to put most weight on semantic dimensions that are predictive of the tag being modeled e.g., "calming, low energy, mellow" music is good for "going to sleep". This demonstration of the dependence between tags indicates the importance of including context when modeling the relationship between semantics and music.

5. CONCLUSIONS

In this paper we have presented the Dirichlet mixture model, a novel approach for improving automatic music tagging by effectively modeling contextual relationships among tags. Starting from the SMN of each song, the DM adds an additional layer to model tag co-occurrences, giving context to the semantic representations derived from

Category	# Tags	Model	P5	P10	MAP
Emotion	35	GMM	0.513	0.506	0.477
		SVM	0.539	0.514	0.481
		DM	**0.561**	**0.535**	**0.489**
Genre	11	GMM	0.367	0.325	**0.355**
		SVM	**0.396**	**0.336**	0.350
		DM	0.360	0.331	0.341
Instrument	14	GMM	0.460	0.431	0.441
		SVM	0.495	0.452	0.455
		DM	**0.506**	**0.458**	**0.463**
Acoustic	25	GMM	0.508	0.501	0.472
		SVM	0.524	0.516	0.471
		DM	**0.564**	**0.546**	**0.496**
Usage	6	GMM	0.253	0.233	0.258
		SVM	0.266	0.226	0.237
		DM	**0.308**	**0.273**	**0.281**
Vocals	6	GMM	0.253	0.240	0.261
		SVM	0.260	0.210	0.235
		DM	**0.287**	**0.267**	**0.278**

Table 2. Retrieval results considering the different word categories for the semantic GMM, and the contextual SVM and DM models.

Context Tag	Semantic Influence		
calming	low energy	tender	slow tempo
hard rock	hard rock	rock	strong
acoustic guitar	slow tempo	tender	acoustic guitar
acoustic texture	low energy	soft rock	light beat
going to sleep	calming	low energy	mellow
emotional	tender	sad	soft rock

Table 3. Examples of the top three semantic influences on contextual tag models.

acoustic content. A tag's affinity with a song is computed as the posterior probability under the tag's DM model. The set of all posterior tag probabilities provides a contextual description of the song.

Experiments reported that modeling context outperforms approaches based on a semantic representation alone, especially considering the top of the ranked retrieval lists. We demonstrate that the DM is an appropriate choice for modeling semantic context by comparison to learning context with an SVM. More specifically, examining the performance across semantic categories, we showed that the DM improves performance for tags that exhibit a high degree of correlation, as well as for noisy tags that are poorly represented by acoustic patterns.

6. ACKNOWLEDGEMENTS

The authors would like to thank N. Rasiwasia for code and suggestions, and N. Orio, and B. McFee for their support.

7. REFERENCES

[1] J.J. Aucouturier, F. Pachet, P. Roy, and A. Beuriv. Signal + context = better classification. In *Proceedings of ISMIR*, 2007.

[2] L. Barrington, M. Yazdani, D. Turnbull, and G. Lanckriet. Combining feature kernels for semantic music retrieval. In *Proceedings of ISMIR*, 2008.

[3] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere. Autotagger: a model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research*, 37(2), June 2008.

[4] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–38, 1977.

[5] M. Hoffman, D. Blei, and P. Cook. Easy as CBA: A simple probabilistic model for tagging music. In *Proceedings of ISMIR*, 2009.

[6] P. Knees, T. Pohle, M. Schedl, and G. Widmer. A music search engine built upon audio-based and web-based similarity measures. In *Proceedings of ACM SIGIR*, 2007.

[7] P. Lamere. Social tagging and music information retrieval. *Journal of New Music Research*, 37(2):101–114, 2008.

[8] C.D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[9] T. Minka. Estimating a Dirichlet distribution. 2009. http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/.

[10] S.R. Ness, A. Theocharis, G. Tzanetakis, and L.G. Martins. Improving automatic music tag annotation using stacked generalization of probabilistic SVM outputs. In *Proceedings of ACM MULTIMEDIA*, 2009.

[11] N. Rasiwasia and N. Vasconcelos. Holistic context modeling using semantic co-occurences. In *Proceedings of IEEE CVPR*, 2009.

[12] M. Sordo, C. Laurier, and O. Celma. Annotating music collections: How content-based similarity helps to propagate labels. In *Proceedings of ISMIR*, 2007.

[13] C.F. Tsai and C. Hung. Automatically annotating images with keywords: a review of image annotation systems. *Recent Patents on Computer Science*, 1, 2008.

[14] D. Turnbull, L. Barrington, and G. Lanckriet. Five approaches to collecting tags for music. In *Proceedings of ISMIR*, 2008.

[15] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE TASLP*, 16(2):467–476, February 2008.

[16] Y.H. Yang, Y.C. Lin, A. Lee, and H. Chen:. Improving musical concept detection by ordinal regression and context fusion. In *Proceedings of ISMIR*, 2009.