

Large Scale Discovery of Seasonal Music From User Data

Cameron Summers and Phillip Popp

Gracenote

Emeryville, CA United States

{csummers@gracenote.com, ppoppp@gracenote.com}

Abstract. The consumption history of online media content such as music and video offers a rich source of data from which to mine information. Trends in this data are of particular interest because they reflect user preferences as well as associated temporal contexts that can be exploited in systems such as recommendation or search. This paper classifies songs associated with a holiday temporal context using a large, real-world dataset of user listening data. Results show strong performance of classification of Christmas music with Gaussian Mixture Models.

Keywords: music, seasonality, machine learning, time series

1 Introduction

Consumption of media content such as music and video often exhibits patterns when associated with a temporal context. Identifying and understanding these contexts can improve the quality of recommendations as shown by [1] and provide useful explanations for the recommendations that are made, improving the user experience [2]. Contexts such as holidays often influence domains beyond music listening, linking music recommendation with other recommendations systems. The importance of holiday contexts in music can be readily observed in industry where flags such as Christmas are often used [3]. However, the task of manually labeling specific content as connected to a holiday is challenging because these connections have a distributed nature - varying by geographic region, language, and time - and expert curation is time intensive and costly. We investigate the feasibility of labeling these connections by classification with user listening data.

Previous research has studied the dynamics and classification of time series signals. In the web search domain, [4] showed that queries could be classified by their change in popularity over time using features in the signal. [5] classified seasonal web search queries using Holt-Winters decomposition on a small data set to improve time-sensitivity in search results. In music listening signals, [6], [7], and [8] show how analysis of temporal dynamics of music listening are useful for recommendations systems and look specifically at seasonality. However, to our knowledge there is no published work that attempts to exploit the temporal analysis of music listening data for automated labeling of holiday music content.

2 Methods and Materials

2.1 Approach

Listen counts of a track will exhibit a differing and detectable pattern around a period of time if it has an association with that period, such as a Christmas track around December 25th. This pattern can be exploited by training a classifier using features of this signal. The features in this study are listening rates of a track i for day j in a window of time localized around the target context

$$R_{ij} = \frac{\sum_{k=1}^U c_{ijk}}{\sum_{l=1}^W \sum_{k=1}^U c_{ilk}} \quad (1)$$

where c is an element of C , and $C \in \mathbb{R}^{T \times W \times U}$ where T is the number of tracks, W is the number of time periods, and U is the number of users. To control for the significant differences in the overall popularity of tracks in a large data set, we normalize the listen counts of each track across the selected periods.

For classification, we chose the Gaussian Mixture Model (GMM) with full covariance matrix because it is fast to train and the listening rates resemble a normal distribution. A GMM is trained using tracks from the target holiday in a training portion of the data set, and classification is performed on the test set using the likelihood of the data given the model.

2.2 Dataset

Number of Records	4,819,992,847
Number of Users	1,648,796
Number of Tracks	13,227,376
Date Range	January 2012 - February 2013

This study uses an internal Gracenote dataset of online radio listening records in North America with some basic statistics of the dataset shown in Table 2.2. Each record of the dataset represents one listen of a track by one user and provides User ID, Date, Time, and Track ID. From the Track ID some associated metadata such as track name and album name is used for keyword search and post-experiment analysis. It is necessary to use a large dataset to get good classification results as shown in section 2.3. Other public datasets similar to Table 2.2 such as “Last.fm Dataset - 1K users” dataset available at <http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/lastfm-1K.html> are too small.

2.3 Experiment - Christmas

We chose Christmas as the target for seasonal music identification because of its popularity and large volume of associated music. We hypothesize that a classifier trained with features in section 2.1 can identify Christmas tracks. We generated an initial set of Christmas tracks by searching for “Christmas” keyword in the

track name and album name - totaling 87,554 Christmas tracks or 0.7% of the entire track population - and maintained a second list of tracks without the keyword. This is not a comprehensive list of Christmas tracks, but is generally free of non-Christmas tracks. Expert curation of a comprehensive set is infeasible with such a large dataset, and using tags from external sources or a more complex text search is error prone.

We chose a consecutive 15 day span centered on December 25th, Christmas, as the listening rate inputs to the classifier. Training and classification (60% train, 40% test) using Gaussian Mixture Models were performed on subsets of the dataset given by tracks with more than some minimum total listens in the whole dataset. To validate performance of the Christmas model, ROC and AUC score were calculated on the test set and are in Figure 1.

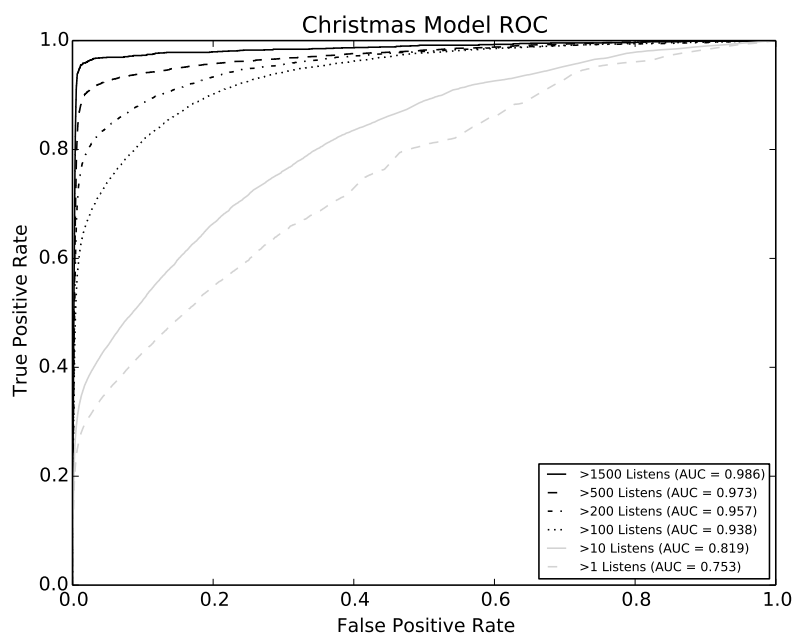


Fig. 1. Test set ROC and AUC of the Christmas classifier on subsets of the dataset thresholded by minimum total listens. The 1,500 listen threshold has 335,678 total tracks with 4,732 Christmas tracks.

3 Discussion

The model performed quite well even though the experiment used an incomplete list of Christmas tracks. At the highest threshold, an inspection of tracks with

high probability according to the Christmas model without the “Christmas” keyword shows that many are other Christmas songs well-known in North America such as “The First Noel” and “Santa Claus Is Coming To Town.” This suggests that the model is not just identifying tracks with the “Christmas” keyword, but would likely accurately classify a more complete list of Christmas tracks.

One notable observation is the change in AUC as the threshold for total minimum listens of track is lowered. Classification suffers when including unpopular tracks. This is likely due to the natural variance in the listen counts of tracks with fewer listens. Normalizing smaller listen counts has a disproportionate effect on computation of listen rates.

The dataset contains only a single year of data, which is a limitation for analyzing seasonal temporal contexts. Multiple years of data could provide better information for classification and show changing listening preferences over time. This is a topic of future work.

4 Conclusion

This study demonstrated on a large, real-world dataset that user listening data could be utilized to detect seasonal music content for Christmas. Classification with a Gaussian Mixture Model showed that the listen rates are sensitive to variance in unpopular tracks and quality results require detection to be performed on a large database of listening records.

References

1. Shin, Dongmin, et al.: Context-aware recommendation by aggregating user context. *Commerce and Enterprise Computing*, 2009. CEC'09. IEEE Conference on. IEEE, 2009.
2. Wang, Weiquan, and Benbasat, I.: Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems* 23.4 (2007): 217-246.
3. The Echo Nest Blog, <http://blog.echonest.com/post/35845347430/christmas-comes-early-to-the-echo-nest>
4. Kulkarni, A., Teevan, J., Svore, K.M., and Dumais, S.T.: Understanding temporal query dynamics. In *Proc. WSDM*, pages 167-176, Hong Kong, China, (2011)
5. M. Shokouhi: Detecting seasonal queries by time-series analysis. In *Proc. SIGIR*, pages 1171-1172, Beijing, China, (2011)
6. Park, Ho, C., and Kahng, M.: Temporal dynamics in music listening behavior: A case study of online music service. *Computer and Information Science (ICIS)*, 2010 IEEE/ACIS 9th International Conference on. IEEE, (2010)
7. Carneiro, Teixeira, M.J.: Towards the discovery of temporal patterns in music listening using Last.fm profiles. *Dissertation*, (2012).
8. Hidasi, Balzs, and Tikk, D.: Context-aware recommendations from implicit data via scalable tensor factorization. *arXiv preprint arXiv:1309.7611* (2013).