Study of Chinese and UK Hit Songs Prediction.

Jianyu Fan[1] and Michael A.Casey[2],

[1]Dartmouth College
[2]Dartmouth College
Jianyu.fan.gr@dartmouth.edu

Abstract. The top 40 chart is a popular resource used by listeners to select and purchase music. Previous work on automatic hit song prediction focused on Western pop music. However, pop songs from different parts of the world exhibit significant differences. We performed experiments on hit song prediction using 40 weeks of data from Chinese and UK pop music charts. We used a set of ten common audio features with a time-weighted linear regression model and a support vector machine model to predict whether a new song will be a top hit or a non-hit. Then we report on the features that performed best for predicting hit songs for both the Chinese and UK pop charts. Our results indicate that Chinese hit song prediction is more accurate than the UK version of the experiment. We conclude that the audio feature characteristics of Chinese hit songs are significantly different from those of UK hit songs. The results of our work can be used to inform how music information retrieval systems are designed for pop music from different musical cultures.

Keywords: Hit song prediction, Chinese, UK

1 Introduction

Pop music in different parts of the world exhibits significantly different musical traits owing to deep-seated cultural preferences. Therefore, MIR systems designed for UK pop music may require different features than those designed for Chinese music. We present a cross cultural case study testing the predictive power of 10 common audio features for music from these two cultures.

A significant problem with cross-cultural comparison is defining tasks that have common meaning. For example, genre or mood classification tasks require labels that would not self-evidently translate to equivalent concepts between cultures [7]. We use hit song prediction in pop music charts. The ranks of songs are measured by the number of sales and radio listeners' short messages vote. Using this task we investigate what makes a hit song in China versus a hit song in the UK. To our knowledge there is no previous work on comparing hit song prediction between cultures.

Hit song prediction has been a recurring, and sometimes contentious [2], topic within music information retrieval [1-4]. The underlying assumption is that "cultural items … have specific, technical features that make them preferred by a majority of people" [2, p. 355]. In [3] it was shown that hit song features vary substantially over time-scales of months and years but remain stable enough over a few weeks to produce better-than-chance predictions. Most of the above studies used a variety of non-linear machine learning methods and, as such; do not easily lend themselves to interpretation of model weights. The current study is concerned with interpretation and feature selection in hit song prediction to compare cultures. Hence, we used linear regression model and we compared these results with using support vector machine model. We address the time evolution of features using time-weighted linear regression classifiers. Our data consisted of official weekly top 40 songs in the year 2012 in the Chinese and UK markets.

The remainder of the paper is organized as follows: Section 2 gives the background to hit song prediction; Section 3 gives details of the dataset used; Section 4 describes the audio features; Section 5 presents methods and results of binary classification; Section 6 gives a comparative cross-cultural analysis of audio features; and we conclude with a summary and discussion in Section 7.

2 Background

A number of systems have been proposed for hit song prediction, which is a binary classification task to predict whether or not a new song will be a hit. Dhanaraj, R [1] used both lyric features and audio features from a corpus of 1700 songs. They used a support vector machine (SVM) with boosting. The reported results were better than random and the results using lyrics features were better than for using audio features alone. Pachet's [2] goal was to classify songs as low, medium or high popularity using a SVM trained on a corpus of 32,000 songs. His results cannot prove their classifiers worked. Ni et al. [3] used the shifting perceptron algorithm, which employs a time-weighted version of the perceptron learning rule, in a corpus of 5947 of the UK top 40 singles over the last 50 years. The goal was to distinguish the top 5 hits from the top 30-40 hits. The accuracy was between 56% and 62%.

In our work, we want to see how hit song prediction varies from different cultures, and we set up our experiments to specifically address that question using 10 common audio features and linear binary classification.

3 Dataset

Top 40 chart data for the year 2012 was collected from the *Official Chart Company*[1] for UK hit songs and *ZhongGuoGeQuPaiHangBang*[2]for Chinese hit songs. We labeled the top 20 songs hits and the bottom 20 not hits. Within these data we also evaluated prediction performance for the highest 5 (1-5) and lowest 5 (36-40) ranked songs. We needed to download Chinese songs one by one to perform audio analysis so we managed to collect 40 weeks of data. Because there is always some gap of weekly data in the Chinese chart, in the 40 weeks of data we have, there are 3 weeks with no data. So we have 37 weeks data on Chinese songs and 40 weeks data from the

[1] http://www.theofficialcharts.com/
[2] http://www.inkui.com

UK chart. Since the chart is Top 40 chart, songs never enter this chart is considered as a new song. There are 347 new Chinese songs in total and 405 new English songs in total.

4 Audio Features

We extracted the following audio features using the EchoNest[3]service to analyze each song: danceability, duration, energy, key, liveness, loudness, mode, speechiness, tempo and time signature. These features most resemble those used in [1] and they are features that globally represent a song. From the EchoNest website, we find the official description of following audio features.

"The danceability feature is a number ranges from 0 to 1, representing how danceable the Echo Nest thinks this song is [1].""The duration feature is the length of the song in seconds [1]." "Energy feature is a number ranges from 0 to 1 representing how energetic the Echo Nest thinks this song is." "Key feature is the signature that The Echo Nest believes the song is in. Key signatures start at C and ascend the chromatic scale. In this case, a key: 1 represents a song in D-flat [1]. " "Loudness feature's description is that: overall loudness of a track in decibels (dB) [1]." "Mode feature is the number representing whether the song is major (0) or minor (1) key [1]." "Time Signature is Time signature of the key; how many beats per measure. [1]." These features are all global features no mature what genre the song belongs to, these features are all meaningful for the song.

In Section 6 we inspect which features were the given weights with higher magnitudes and did the hit song prediction using different subsets of features based on selecting the higher-weighted feature terms. We obtained different feature weightings for Chinese hit songs prediction and UK hit songs prediction, as discussed below.

5 Machine Learning

To predict whether a song will be ranked higher or lower, and to analyze the feature weights of predicting UK songs and Chinese songs, we used a time weighted linear regression and compared results with a support vector machine model. We used time-weighted linear regression (TWLR) and support vector machine (SVM) to predict whether songs newly entering the char in the coming two weeks will be hit songs.

5.1 Time Weighted Linear Regression (TWLR)

To account for feature variation in time we give more weight to the training data that is closer in time to the test data so that the model prediction results are more affected

[3] http://the.echonest.com/

by more recent data and less by data in the more distant past. Locally (time) weighted linear regression [5] is defined as the following: For a given training set,

$$D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\} \quad \textbf{(1)}$$

and for a given test example, x, we fit θ by minimizing:

$$E(\theta) = \frac{1}{2}\sum_{i=1}^{m} w^i(x)\left(y^{(i)} - \theta^T x^{(i)}\right)^2 \quad \textbf{(2)}$$

where θ^T is the linear regression weight matrix and $w^i(x)$is a time weight dependent up on the temporal distance between x and $x^{(i)}$. Then:

$$w^i(x) = exp(-1 * |t(i) - t^*(i)|) \quad \textbf{(3)}$$

with $t(i)$ the time of the training data ($t(i)$ is 1 for the first week's data) and $t^*(i)$ the time of the test data. To avoid numerical problems we scaled the weights w^i(x)to add to 1 for each x. If $\mathrm{w}^{\mathrm{i}}(\mathrm{x})$ is small then the error terms $\left(\mathrm{y}^{(\mathrm{i})} - \theta^{\mathrm{T}}\mathrm{x}^{(\mathrm{i})}\right)^2$are negligible. If $\mathrm{w}^{\mathrm{i}}(\mathrm{x})$ is large the algorithm adjusts the weights to reduce the error. We can compute $\theta(x)$ using:

$$\theta(x) = (X^T W X)^{-1} X W y \quad \textbf{(4)}$$

with W a diagonal matrix of the temporal weights.

We used a shifting four-week window on the data to perform training and prediction, with the first three weeks of data in each window used for training and new songs in last week used for testing. The window was advanced by one week and the process repeated. We also used a shifting five-week window on the data while the first three weeks of data is used for training and the last week is used for testing.

We defined hit songs as those with rank 1-20 and non-hit songs as those with rank 21-40, thereby yielding an equal chance of randomly assigning the correct label. For each window, we count the number of songs that were accurately predicted among new songs (Songs haven't entered Top40). Each week there are 7 new songs on average. In addition, we count the number of top 5 songs among new songs (ranks 1-5) that are predicted as hit songs and the number of bottom 5 songs among new songs (ranks 36-40) predicted as non-hit songs.

Table 1 shows the results of predicting new UK and Chinese hit songs for the TWLR and SVM model. The results of TWLR indicate that performance of predicting Chinese songs was significantly above the baseline (50%). (Err = 41.58%; p-value = 0.03). But TWLR doesn't work very well for predicting new UK hit songs. (Err = 52.10%; p-value = 0.29)

As for SVM, we chose RBF kernel and gamma is equal 100. SVM model works great for predicting both new Chinese songs (err = 39.25%; p-value = 0.011) and UK hit songs (err = 42.30%; p-value = 0.04). However, the result of predicting UK new songs which are from top1-5 and top 36-40 is not very significant (err = 44.32%; p-value = 0.263). The overall result of predicting 2nd week's data is less accurate than those of predicting 1st week's data

Table 1. Error Rate for New Songs Prediction

Time	Method	Data	Error rate (40 songs)	P Value	Error Rate (1-5vs. 36-40)	P value(1-5vs.36-40)
Week1	TWLR	UK	52.10%	0.290	46.39%	0.287
		Chinese	41.58%	0.030	22.22%	0.004
	SVM	UK	42.30%	0.004	44.32%	0.263
		Chinese	39.25%	0.001	29.62%	0.038
Week2	TWLR	UK	46.92%	0.882	43.29%	0.133
		Chinese	56.84%	0.330	44.44%	0.806
	SVM	UK	41.22%	0.005	43.29%	0.152
		Chinese	44.50%	0.147	29.62%	0.021

5.2 Result Analysis

There is no overlapping between training and testing sets. By comparing the results of using TWLR and SVM, we can see that SVM performed better with yielding significant results relative to the baseline. Thus, whilst a linear hyper-plane in the feature space cannot powerfully separate the categories there are linear categorical tendencies in the feature space. The SVM is able to project the features into a kernel space to perform the separation so it is difficult to interpret which features contribute most to the classification. However, TWLR gives us detailed information about the weights of features so that we explore the differences of feature between UK hit song prediction and Chinese hit song prediction.

Table 1 shows that model prediction is better for top 5 and bottom 5 songs, except for UK Hit Songs while using SVM. We propose that this is because the top and bottom ranked songs are more likely to exhibit the feature traits learned by the classifier. Also, for these two groups, there are fewer changes in position week-to-week relative to other songs. Therefore the time-weighted model could more easily predict songs based on the earlier week's data.

In addition, Chinese song prediction was significantly better when we only considered new songs from top and bottom 5 songs than considering from all Top 40. We analyze the social background and the results of the test. We can know that whether the song is rap and whether it is recorded in concert or in studio will affect the rank of the song in Chinese chart greatly. Again, we interpret this to mean that those song positions are most likely to exhibit the traits of the hit/non-hit categories making them stand out from the songs at middle ranks with less certain labels.

5.3 Time to Become a Top10 Hit

To check whether it is better to predict the next week's data and the second next week's data, we counted the number of weeks hit songs take before becoming a top10 hit. Figure 1 and Figure2show the distributions of number of weeks of UK and Chinese hit songs.

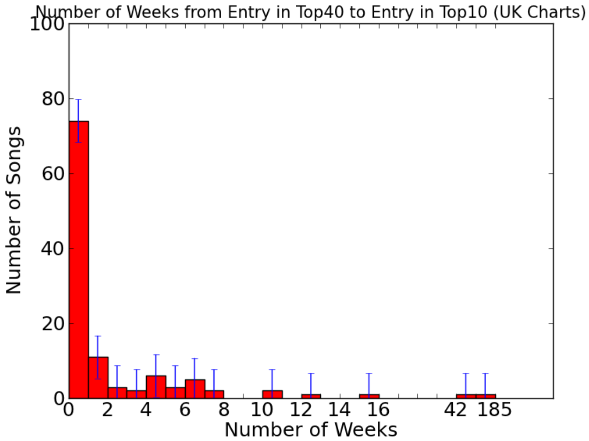

Fig. 1. Number of Weeks before Becoming Top10 Hits (UK Charts)

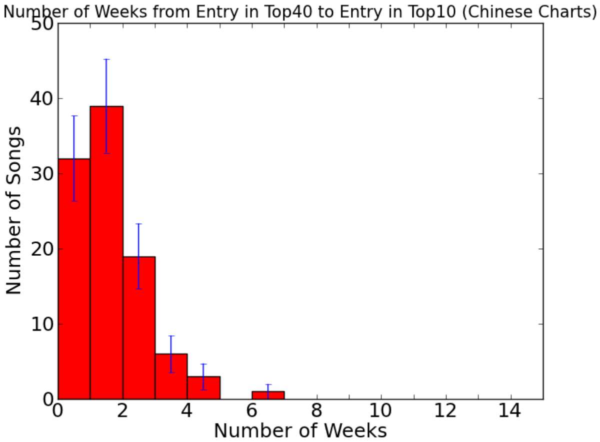

Fig. 2. Number of Weeks before Becoming Top10 Hits (Chinese Charts)

The distribution in above charts shows that over 2/3 of top10 hits are brand new songs in UK chart. While for Chinese chart, over 2/3 of top10 hits take zero week or one week to get in to top10.

Figure 3and Figure 4 show the distributions of number of weeks of songs stay in top 10.

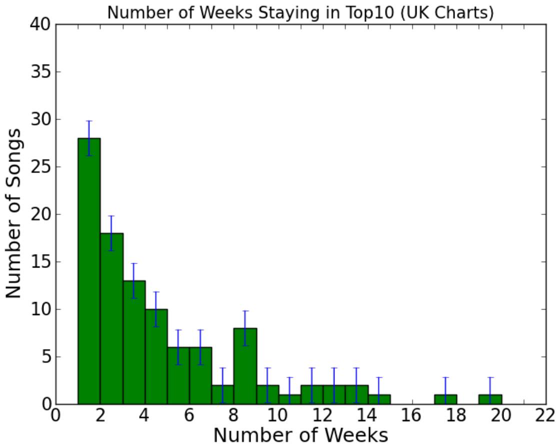

Fig. 3. Number of Weeks Staying in Top10 Hits (UK Charts)

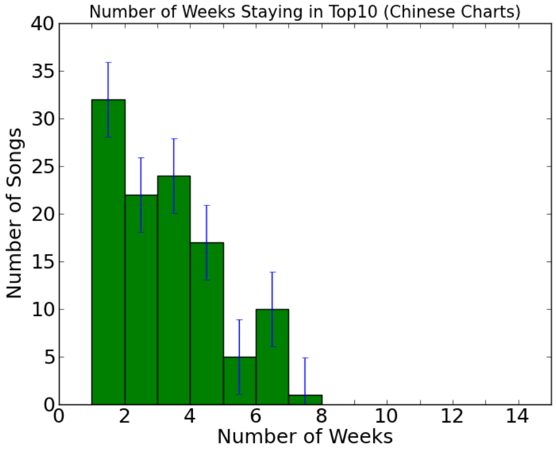

Fig. 4. Number of Weeks Staying in Top10 Hits (Chinese Charts)

6 Analyses of Features

We inspected the linear regression weight in both the Chinese hit songs predicting model and UK hit songs predicting model to discover which features are more important. In addition, we did both UK hit song prediction and Chinese hit song prediction using increasing number of features.

6.1 Effects on Results Using Different Features

6.1.1 Feature Weights

The θ^Tvector gives us information about the weight of each feature. Figure 5 and Figure 6 display the result of models for UK hit songs and Chinese hit songs. The x axis represents features. Table 2 shows the corresponding features.

Table 2. Feature Index and Corresponding Feature Name

Feature Index	Feature	Abbreviation
1	danceability	dan
2	duration	dur
3	energy	eng
4	key	key
5	liveness	liv
6	loudness	lou
7	mode	mde
8	speechiness	sch
9	tempo	tep
10	time signature	tsig

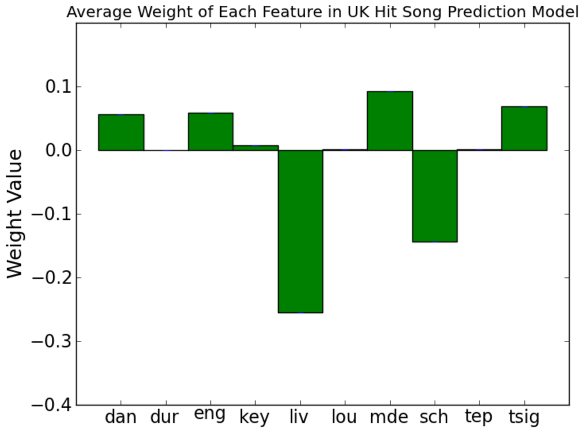

Fig. 5. Average Weight of Each Feature in UK Hit Song Predicting Model

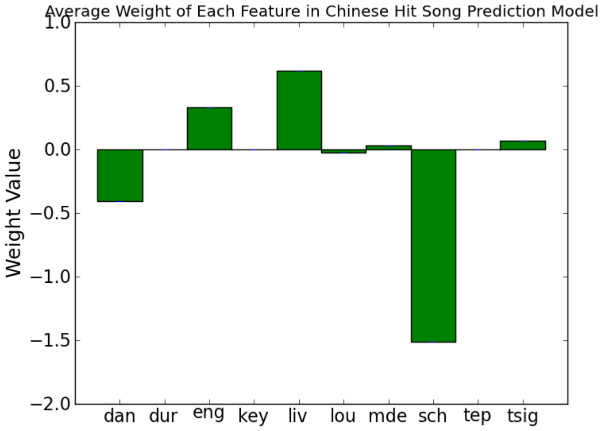

Fig. 6. Average Weight of Each Feature in Chinese Hit Song Predicting Model

Figure 5 and Figure 6 show that danceability, energy, liveness, mode, speechiness, and time signature are more important when predicting UK hit songs; danceability, energy, liveness and speechiness are more important when predicting Chinese hit songs.

6.1.2 Increasing Features in UK Hit Songs Prediction Weight

The value of elements in θ^T shows that the features are ranked as follows (listed from highest weight to lowest weight): liveness, speechiness, mode, time signature, energy, danceability, key, loudness, tempo, and duration. We started with using the three most important features to do prediction and then we increased the number of features by adding the next important one until all of them were used.

Figure 7 shows the results. The x axis in Figure 7 represents the number of features. When it's3, it means we used the three most weighted features to do the prediction which are liveness, speechiness, mode. When it's 4, it means we used the four most weighted features to do the prediction which are liveness, speechiness, mode, time signature.

The error bar shows the error rates of distinguishing 1-5 vs. 36-40.

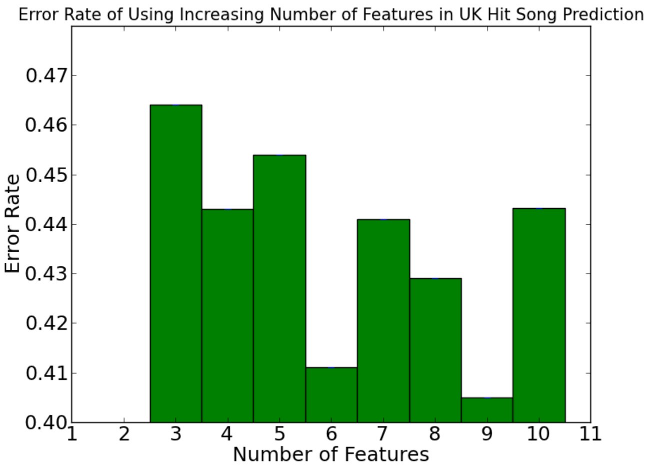

Fig. 7. Result of Using Increasing Number of Features in UK Hit Song Prediction (order in weight, see 6.1.2.text)

From Figure 7, we can see that the error rate generally decreases with the increase of number of features.

6.1.3 Increasing Features in Chinese Hit Songs Prediction Weight

As for Chinese hit songs prediction, features with higher weight to lower weight are ranked as following: speechiness, danceability, liveness, energy, time signature, mode, key, duration, tempo, and loudness. We did the same experiments as in 6.1.2.

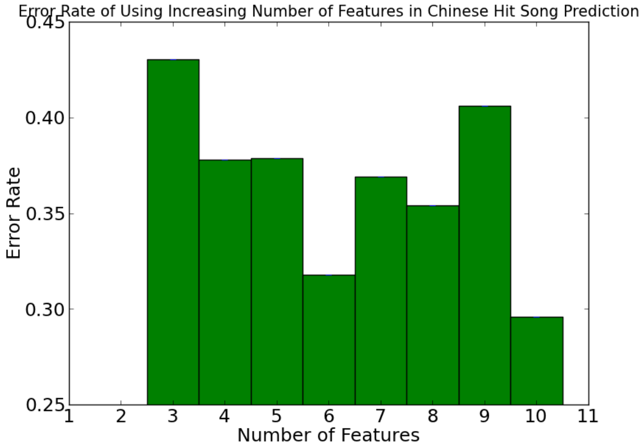

Fig. 8. Results of Using Increasing Number of Features in Chinese Hit Song Prediction (order in weight, see 6.1.3. text)

The x axis in Figure 8 represents the number of features. When it's 3, it means we used the three most weighted features to do the prediction which are speechiness, danceability, liveness. When it’s 4, it means we used the four most weighted features to do the prediction which are speechiness, danceability, liveness, energy. From Figure 8, we can see that the error rate generally decreases gradually with the increase of number of features.

6.2 Features Comparisons between Chinese Hit Songs and UK Hit Songs

We compared features of UK hit songs with those in Chinese hit songs. The Danceability, Energy, Speechiness and Tempo features vary greatly. In the following figures, the red line indicates the UK top 5 songs and the blue line indicates the Chinese top 5 songs.

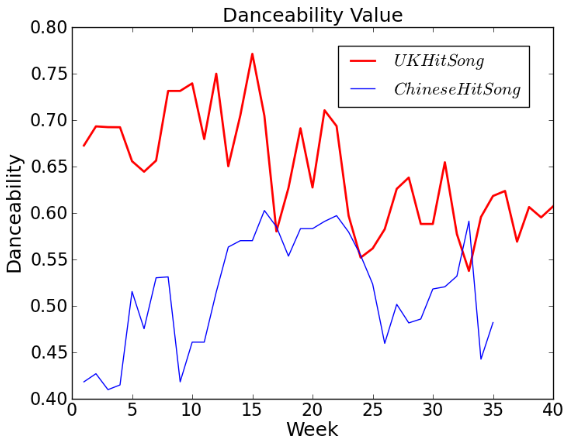

Fig. 9. Danceabilityof UK and Chinese Top 5Hit Songs

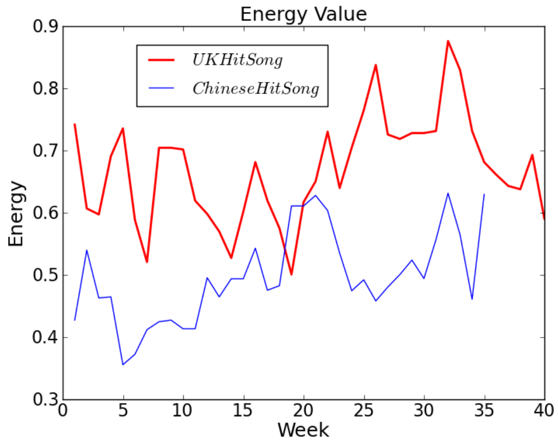

Fig. 10. Energy of UK and Chinese Top 5Hit Songs

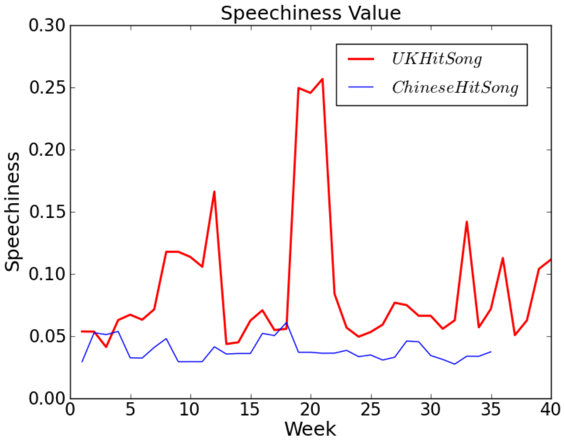

Fig. 11. Speechiness of UK and Chinese Top 5Hit Songs Tempo

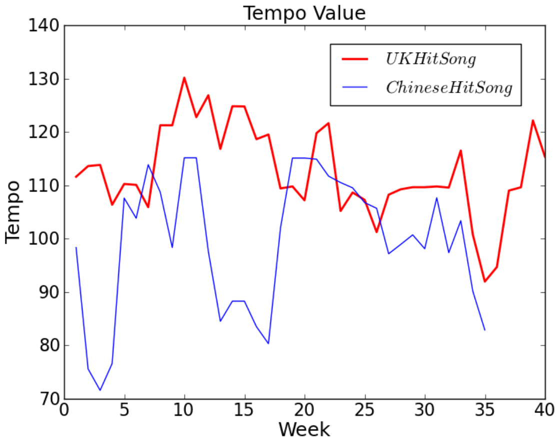

Fig. 12. Tempo of UK and Chinese Top 5Hit Songs

From Figure 9, Figure 10, Figure 11 and Figure 12, it is obvious to see that values of danceability, energy, tempo and speechiness of the UK top 5 songs are higher than those of the Chinese top 5 songs. We can generalize that Chinese hit songs are more melodic and less energetic and much less songs are suitable for dance parties.

7 Conclusion and Future Works

We conducted Chinese and UK hit song prediction and compared between them. We used a time- weighted model which has same concept used in Yizhao Ni and Matt Mcvicar's model [3]. The test we conducted differed from [1] [2] [3].Our results show that the prediction result is promising. It proves that the hit songs prediction is doable.

It is interesting that a simple model as linear regression works on the problem without considering the hyper plane. Though the results of TWLR are not as good as SVM model, TWLR is an easier to analyze the differences between characteristics of Chinese hit songs and UK hit songs. We pointed out the features that are more significant for each prediction. The error rates are generally getting lower when using increasing number of features. Our test indicates that it is possible to predict trending tracks well during local time periods under different cultural backgrounds. In addition, the feature comparison shows the obvious differences between Chinese hit songs and UK hit songs, indicating that Chinese hit songs are more melodic, slower, and less energetic. Chinese pop music has the reflection of Chinese's traditional music which is much more calm and melodic. The rock and roll music was developed in Western countries much earlier than in China which also affect the characteristic of Chinese and UK hit songs. We believe that more experimentation should be done using different features, using different models of prediction and also conducting more comparisons.

References

1. Dhanaraj, R., and B. Logan: Automatic prediction of hit songs. In: Proceedings of the International Conference on Music Information Retrieval. (2005)
2. Pachet, P. Roy: Hit song science is not yet a science. In: Proceedings of ISMIR, 355–360, (2008)
3. Ni, Y., Santos-Rodriguez, R., Mcvicar, M., & De Bie, T.: The T Hit Song Science Once Again a Science? In: 4th International Workshop on Machine Learning and Music Learning from Musical Structure, (2011)
4. Pachet: Music Data Mining, Editorial Tzanetakis, Ogihara Tao, (2011)
5. W. Cleveland and S. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting" J. Amer. Statist. Assoc., vol. 83, no. 403, pp. 596–610, Sept. (1988)
6. Cortes, C. and Vapnik, V. Support vector machines. Machine Learning, 20, 273-297. (1995)
7. Yang, Y. & Hu, X. Cross-cultural Music Mood Classification: A Comparison on English and Chinese Songs. In: Proceedings of the 13th International Society for Music Information Retrieval Conference. Porto, Portugal. October 8-12, (2012)