Lawrence Atherton

Remix Novelty Ranker

for Selected Trance Music

Introduction and Goal

The original goal of this project was to build a classifier that, when presented with a remix of a song and the corresponding original mix, would classify the remix as "interesting" or "not interesting," thus directing the user as to whether it would be worth their time to try listening to or buying the remix if they enjoyed the original (or, conversely, if the two songs are similar enough that the remix is not worth their time).

My motivation for this project was my experience with the genre of Trance music, where a few cultural and economic factors work toward producing remixes that are overwhelmingly bland rehashes of the original, but occasionally are very well done and may even surpass the original. Remix artists are usually paid a flat fee instead of a commission for their work, and artists tend to save their best ideas for their own work. However, sometimes an artist really falls in love with a song, and does an excellent job remixing it. This is why I thought it would be useful to try to classify all remixes into those two disjoint sets.

Initially, I tried using k-means to cluster the data, and attempted to find a k such that 10-20% of the remixes would map to a different cluster than their original; I would have labeled those which map to the same cluster as "not interesting" and those which map to a different cluster as "interesting." However, my data had a high enough dimension and a large enough spread that this was impossible for any k. The best result I got was labeling around 45% of the data as interesting with k=2, which was a little un-nuanced for my liking.

Because of these initial results, I decided instead that I would rank the remixes according to how different they were when compared to their original mix, instead of making a binary classifier. I decided to call this new metric "novelty," a measure of how unique a remix is in the small scope of a single release.

For my data set, I used the first 110 relevant[1] releases of a particular Trance label called Anjunabeats, which collectively have 209 remixes (most of which are also Trance).

A Description of the Algorithm

To the aim of ranking remixes according to their novelty, I developed several independent ranking algorithms. After I saw that many of the algorithms produced similar rankings, especially among the top 20 and bottom 20, I decided that I would treat the individual ranking algorithms as weak learners and work within the realm of ensemble methods. Since each of the algorithms may have vastly different units, and since I don't trust any one algorithm's output too much, I drop all units and only work with rankings outside of the atomic level of a single ranking algorithm.

The decision to drop any units the moment I produce a ranking gives me an interesting property – I can effectively "downweight" any particular algorithm by combining it with another; this action (which I call "ensembling") produces another unit-less ranking and thus can be performed many times recursively. Put another way: say that I have rankings A, B, and C, and I want my final ranking to reflect 50% of C and 25% of A and B. I can call `ensemble(ensemble(A,B),C)` to achieve this goal.

For features, I found a histogram of pitches (shifted by key so that the key of the song is irrelevant), a histogram of 16th notes[2] for one measure averaged over all the measures, and the

---

[1] I ignored releases that had two original mixes, without a remix for either of the originals, and releases that had two versions of the same song, one with and one without vocals. (However, one of the latter type escaped me and actually became an important part of my data set.)

average of the mel-frequency cepstral coefficients, which capture the spectrum of power across frequency[3]. These three sets of histograms collectively measure analogues of melody, rhythm, and timbre, though they necessarily throw away much global temporal information so that my feature space doesn't explode in size beyond the number of data points I have[4]. To preprocess the data, I scaled the pitch and rhythm histograms by 10 and the first mel-frequency cepstral coefficient by 0.1, so that all the features were on the order of 10^0.

Here is a list of ranking algorithms I used, along with a small explanation of what each does and how I used it in the final ensemble ranking:

Naïve distance to original: I measured the Euclidean distance[5] between each remix point and its original, then used these distances to rank the remixes (with larger distance as better). I did this both in the original feature space and in one with 0 mean and unit variance, then averaged the two rankings. This way, I kept some information about the original magnitudes of the features, but the ranking wasn't too skewed by any one feature with a much larger spread than the rest.

Number closer than original: Very similar to the naïve distance method, but substitutes Euclidean distance with the number of points that are closer to the remix than the original is, according to Euclidean distance. This allows for more of a sense of local density / nearest neighbors than the naïve Euclidean distance allows.

"Half" K-Means: The ranking produced if you take the Euclidean distance between the centroid for each pair of points doesn't make a good ranking, since several pairs of points have exactly the same distance (and several even have a distance of 0.0). Therefore, this ranking metric takes the Euclidean distance between a remix point and its original's centroid. I ensemble the rankings for k=5 and k=10; k=20 occasionally crashed so I took it out.

"Half" Gaussian Mixture Models: Like "Half" K-means, but with Gaussian mixture models. I use k=10 and k=50, since GMMs didn't have the crashing problem that k-means had for high k.

PCA Ensemble: One attempt at dimensionality reduction, since I feared that a dimensionality of ~40 was too high for only ~200 remix points. This is a higher order function that performs PCA on the data points, then passes them to one of the above functions. In the final algorithm, I do PCA to dimensions 1, 2, 3, 5, and 10, then ensemble all the resulting rankings.

Feature Group Ensemble: Another attempt at dimensionality reduction. This is a higher order function that splits the features into the pitch, rhythm, and timbre features, then passes these features to any of the algorithms above (including the PCA ensemble function). Then, it ensembles the three resulting rankings together.

---

[2] The 16th note is the tatum for all of the songs.
[3] I used aubio for event detection and MFCCs, and the Melodia Vamp plugin for pitch.
[4] I attempted to also use the variances of each mel-frequency cepstral coefficient in order to preserve a little more global information from each song, but I found that in practice, these features really just injected noise into my algorithms. Because of this, I ultimately left them out.
[5] I also tried using cosine distance, but found that it threw out too much information from magnitude in order to be a useful distance metric for this particular project.

The Final Algorithm: I name the first four functions described above the "basic functions." I call each of the basic functions to obtain four rankings. Then, I use PCA before calling each of the basic functions to obtain four more rankings. Then, I use the Feature Group function on the basic functions to obtain four more rankings. Finally, I use the Feature Group function on the PCA function, which in turn is used on the basic functions, for four final rankings. I ensemble each group of four and then ensemble the four groups together. All of the above constitutes one run of the final algorithm. For the statistics in this writeup, I ensembled 20 such runs together to reduce the impact of any instability in the algorithm.

Kernel Density Estimation / A Second Metric – "Strangeness": I used kernel density estimation in order to get an idea of how closely related any one song is to the rest of the label's catalogue. I decided to call this metric "strangeness." Note that as a concept it's slightly different than novelty: novelty measures uniqueness within the small scope of one single release, whereas strangeness measures uniqueness within the scope of the entire label. Since it attempts to emulate a probability distribution function, KDE outputs higher values for denser regions, so the ranking is from lowest value (least dense; strangest) to highest value.

Results

Since I don't have labeled data, it was difficult to get an idea of just how well my algorithms worked. However, before I ran the algorithms, I identified a few remixes that I thought might be very novel or not novel. I decided that remixes that appeared on the label's yearly compilations (which mostly feature original mixes) might have more of a chance of being novel because appearing on a compilation at all is significant. I found a few remixes that are almost exactly the same waveform as their original mix as candidates for being very un-novel. I've listed these songs and my reasoning for identifying them in more detail in Appendix 1.

Of the remixes that were featured on compilations, four[6] of eighteen (22%) were in the top 10% of the novelty ranking. The five remixes I identified as potentially very un-novel occupied five of the last six spots in the ranking. The sixth spot is occupied by a song that is not so much a remix as a mix by the original artist with small cosmetic changes.

What's more interesting is the remix occupying the seventh to last spot, "Nova (Daniel Kandi vs. Kris O'Neil Remix)." The remix brings on a new artist[7], has a drastically different tempo[8], and could be argued to have changed the subgenre of the song (from mid-2000s Anthem Trance to late-2000s Progressive Trance). Still, the remix reuses the main melody, uses synths with very similar timbres, and has a similar groove to the original mix. **I would consider this a case against pure metadata-based machine learning on music** – the artist is different, the bpm is different, the subgenre is different, but the more essential parts of the song (melody, rhythm, timbre) remain relatively unchanged, and the remix is the worst-ranked out of all the remixes that are not just a version of their original with negligible changes[9].

---

[6] I Kill for You (Probspot Remix) [0th], Eighties (Ozgur Can Remix) [5th], Amsterdam (Smith & Pledger Remix) [11th], and Breaking Ties (Jaytech & James Grant Remix) [15th]

[7] Daniel Kandi is the original artist, but Kris O'Neil is new to the remix.

[8] The two songs differ in tempo by 4 bpm, while most remixers keep the original tempo. It may not sound like much, but 4 bpm can completely change a song's genre in electronic dance music.

[9] Note that being un-novel is not necessarily bad; it depends on whether a listener wants more of the same or a fresh take on the original. I happen to enjoy both mixes of Nova.

Another interesting observation is that Above & Beyond have quite a few remixes in the top 10% by novelty (12 of the top 21, or 57%). Above & Beyond are the heads of the label, and they usually pay for their songs to have on the order of 4 or 5 remixes instead of the usual 1 or 2. Because they pay for many remixes, it makes sense that they have a higher concentration everywhere in the ranking, compared to most artists. However, the remixes of their songs have an average ranking* of 41.09 (compared to the average of 50 for all remixes). Thus, it's clear that not only do Above & Beyond pay for more remixes for their songs, but the remixes they pay for are also usually of a higher quality.

Novelty may not be the best predictor of whether a remix is used on a compilation. As an example, 16 Bit Lolitas remixed "On a Good Day" twice. One of these (the "Downbeat Remix") is in the top 10% by novelty at rank 17, but the other one (rank 66) is the one featured in a compilation. Listening to the two remixes, I would subjectively say that the less novel one fits better with the rest of the songs on its compilation. Also, the less novel one is stranger – ranked 75th by strangeness compared to 159th for the "Downbeat Remix."

Let's take a closer look at the strangeness metric. Since it's a metric on all the songs in the label, we can say not only what a remix's strangeness ranking was but also compare it to its original's strangeness. Out of all remixes featured on compilations, 12 of 18 (67%) are stranger than their original, whereas across the entire label, only 53.6% of remixes are stranger than their original. Compilation remixes are about 18.5% higher in the strangeness ranking than their original, whereas an average remix is only 4.1% higher. The average strangeness ranking* for a compilation remix is 37.87, compared to 49.39 for all remixes and 50 for all songs. For the same songs, the difference using the novelty metric is much less stark. The average novelty ranking* for a compilation remix is 47.36, compared to 50 for all remixes. Clearly, strangeness is a much better predictor of whether a remix will be in a compilation: compilation remixes are 11.5 percentiles more strange than the average, whereas they are only 2.6 percentiles more novel.

However, novelty and strangeness are not the be-all, end-all when it comes to whether a remix will be featured on a compilation or become popular. Consider the three remixes "I Kill for You (Probspot Remix)," "Eighties (Özgür Can Remix)," and "Oceanic (Satoshi Fumi's Unduation Remix)." All three are very novel (ranks 0, 5, and 7), and all three are very strange (ranks 5, 6, and 8). However, while the first two are featured on compilations, the third is not (its original mix is used instead). Furthermore, from looking at the tracklists from Above & Beyond's radio show, I can tell that Oceanic's original mix and two of its more conventional remixes (by Sean Tyas and Super & Tab) were all played several times over the years, but Satoshi Fumi's remix was never played once. From listening to the music, it might be that Above & Beyond never played Satoshi Fumi's remix because it is very much a House track, whereas the rest of the remixes are Trance. However, the 16 Bit Lolitas remix I mentioned earlier is also a House track, and it was featured on a compilation. I believe that the reason Above & Beyond didn't play the Satoshi Fumi remix more frequently is that it simply came out at an inopportune time; Oceanic was released in 2007, but Above & Beyond didn't really start advocating House music until two years later in 2009 with the release of their first House compilation, Anjunadeep:01, which is the compilation that 16 Bit Lolitas' remix was featured on.

In summary, strangeness is a better predictor than novelty of whether a remix will be featured on a compilation. Strangeness is a metric of how interesting a song is in relation to the entire label, and novelty is only a metric of how interesting a song is in relation to the rest of the

---

* Normalized to [0, 100].

songs on its single, so novelty can tell you what remixes to prioritize listening to given a list of original mixes that you enjoy, but strangeness can tell you what songs are most interesting out of the entire label. Generally, more interesting songs are put on compilations. Both metrics work well for what they are defined to do. However, neither metric is a perfect predictor of popularity or success, simply because a lot comes down to what the heads of the label decide to push.

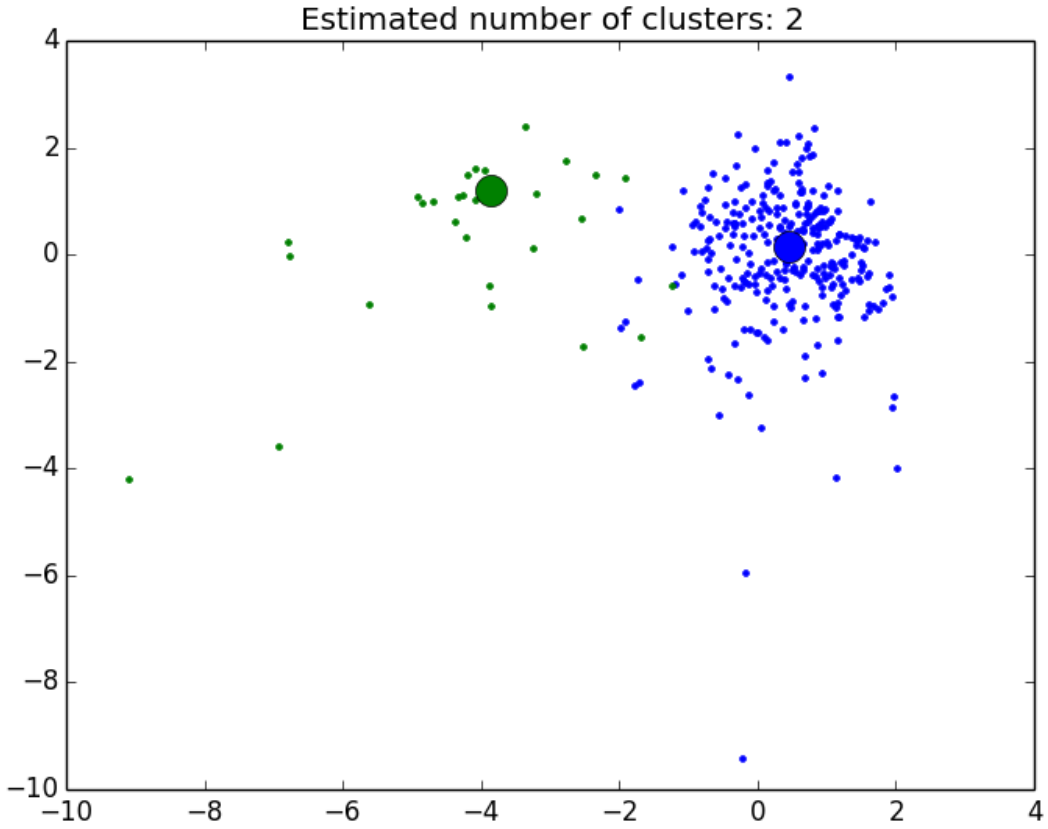
Future Directions

I think the most important direction that audio machine learning research can take is getting better features. My features necessarily threw away most temporal information, keeping only rhythm within one bar, because the number of data points I had was so small. Better features would keep track of structure in some way; for example, interval histograms for pitches, features encoding phrasing for rhythm, spectral flux, features for the rate of information in the raw audio file over time, etc. Key is also important; for example, artists who write music with a heavy bassline often choose to write their songs in E, since the lowest note we can hear is an E, and their songs are often played in clubs using high-end speakers capable of producing that note.

A possible feature space for remix comparison specifically is information divergence / Kullback-Leibler divergence; however, this would be hard to tune, as a remix of an original song doesn't necessarily share any aspect of structure with the original song, in the way that two performances of the same work do.

Appendix 0. Fun with Mean Shift

When you run mean shift on the pitch histograms, you get 2 clusters, corresponding roughly to songs in a minor key and songs in a major key. (I verified this by listening to clips of each of the songs in the latter category, which is smaller and generally agreed with all my expectations from prior experience with the label.) The major cluster has a similar size to the minor cluster, despite being less dense, because it represents more modes (Ionian, Mixolydian, and Lydian), whereas most of the songs in the minor cluster are Aeolian, with only a few Dorian songs.



Appendix 1.

Remixes that might be novel:

The label releases compilations of their songs on an approximately yearly basis. The first compilation is mostly comprised of remixes, as the label was just starting out, but the rest of the compilations are comprised mostly of original mixes. Because of this, I assumed that there might be something significant about remixes that manage to appear on compilations (starting with the second). Here's the list of all remixes in my data set that appear on compilations:

I Kill For You (Probspot Remix)
Kallocain (Robert Nickson Remix)
No One On Earth (Gabriel & Dresden Club Mix)
Eighties (Ozgur Can Remix)
Air For Life (Airwave Remix)
Sirens of the Sea (Kyau & Albert Remix)
First Aid (Perry O'Neil Remix)
Amsterdam (Smith & Pledger Remix)
Can't Sleep (Maori Remix)
Needs to Feel (Wippenberg Remix)
One Night in Tokyo (DJ Shah's Savannah Mix)
Cold Front (Bart Claessen Remix)
These Shoulders (Club Mix)
Mount Everest (Dennis Shepard Remix)
Aurora (Sunny Lax Remix)
Eclipse (Mat Zo Remix)
Breaking Ties (Jaytech & James Grant Remix)
On A Good Day (16 Bit Lolitas Remix)

Remixes that might rank very low in novelty:

As I mentioned before, I accidentally let one release into the data set where the remixes are just versions of the original with more or less vocals, or phrases slightly moved around (Won't Sleep Tonight). Another remix (These Shoulders) is exactly the same as the original, with a few phrases taken out. A third remix (Sunrise) is just slightly edited in the intro and outro in order to fit into one of the compilations (and is even credited as the original mix in the compilation).

Won't Sleep Tonight (Original Dub Mix) / (Moody Vocal Mix) / (Moody Dub Mix)
These Shoulders (Club Mix)
Sunrise (Volume 6 Edit)

There are many more remixes in the first category because I don't have a great idea of what remixes should be considered novel, so I'm effectively casting a wide net and looking systematically at all the remixes that I think have a good chance of being novel; in contrast, I have a pretty firm idea of the least novel a song could be (i.e. almost exactly the same), so I'm hoping my algorithm will detect those remixes.

Appendix 2. Libraries Used

http://aubio.org/
http://mtg.upf.edu/technologies/melodia
http://www.numpy.org/
http://www.scipy.org/
http://scikit-learn.org/
http://stackoverflow.com/questions/13224362/pca-analysis-with-python (with small edits so the code actually runs)