Josh Dinges

Data Mining

December 16, 2011

KDD Cup Challenge

For the last project, I was tasked with accessing the data used in Yahoo’s KDD Cup Challenge, mining it, and building a classifier. To be completely honest with you, my classifier, if I can get one, won’t be good. Unfortunately, I forgot about the portfolio assignments until Thursday and today (the 16th) has been a glorious 16 hour battle to hammer through all of those. Needless to say, I haven’t been able to devote much time or brainpower to this assignment. This is unfortunate, because of the time I have spent with the assignment; I do find the concept and the data interesting. Regardless of my intentions, however, I will present my findings in this paper, no matter how limited they may be.

The KDD Cup Challenge was hosted by Yahoo to challenge amateur data miners to predict musical preferences based on data that Yahoo Music has collected. The data is separated into several musical items: tracks, albums, artists, and genres. Each of the items is based on an internal hierarchy where a track belongs to an album, an album belongs to an artist, and artists belong to genres. Yahoo provided participants with the data in the form of text files on album, artist, genre, track, summary statistics, training, and validation.

My theory for a classifier goes a little like the following. Each user has a particular set of tastes when it comes to music. Those tastes are can be better described by listing the genres that they enjoy. My theory is that users will rate similar combinations of genres based on their preferences higher than those that do not match their tastes. Another way of thinking of this is to compare a person’s taste in, say movies. Most critics will agree that the general categories for movies are as follows: action adventure, drama, comedy, suspense/thriller. Now, lumping all of the films that have ever been released into those four categories isn’t going to be very representative of their tastes. So, if we start introducing more genres and more sub-genres, I think we’ll start to get a better picture of what’s going on. If we introduce a romance genre and then combine it with comedy, we may have found something special/appealing – the romantic comedy. For whatever reason, these movies make money and are released every year. Why? Because they appeal to a certain type of moviegoer’s tastes. In fact, it can be argued that people who like one or two romantic comedies will often like several more, but perhaps not action movies. Now, take this situation and apply it to music. We have certain users who might, say, rate a rock, alternative, indie mix highly, but a country, western, electronica mix low. It’s highly likely, that given a user’s taste for say the Foo Fighters, they’ll also like, say, Tenacious D, but not Garth Brooks.

To test my theory the first thing I did was import all of the .txt files into a database. Unfortunately, this is where I ran into difficulty and couldn’t progress much farther due to time constraints. That being said, had I been able to continue, I would have done the following: start pulling from the database as a means to explore my data by using some interesting queries. Next, I would have opened the data up in WEKA and see if I could generate any classifiers from there, however, I’d be much more interested to see what rules WEKA could find from looking at the data set.

Regardless of what I meant to do, I admit that I did not accomplish much on this assignment. I hope you don’t take that as an indicator of how I felt about the class as a whole because Data Mining was one of the most interesting and useful classes I’ve taken during my time at Mines.