5004-deep-content-based-music-recommendation:

Abstract:

Eine automatischer Musik Vorschlag wurde in den letzten Jahren ein immer relevanteres Problem, da der Großteil der Musik nur noch online verkauft und konsumiert wird. Die meisten Systeme filterten Musik nach einer gleichen Meinung. Allerdings hat dieser Ansatz das „Cold Start“ Problem, es scheitert dort wo keine Daten vorhanden sind, wodurch es sich nicht eignet neue oder unbekannte Lieder vorzuschlagen. Dieses Paper befasst sich mit der Nutzung eines Eigenvektors für Vorschläge und trifft aussagen über Eigenvektoren von Musik, wenn sie nicht aus den vorhanden Daten geschöpft werden kann. Wir vergleichen den traditionellen Ansatz „Bag-of-Words“ mit dem der konvolutionellen Neuralen Netzwerken und bewerten die Vorhersagen-Qualität und -Quantität der Millionen Song Datenbank. Wir zeigen das das benutzen von Eigenvektoren eine sinnvolle Empfehlung erzeugen, trotz der Tatsache, dass es eine große semantische Lücke zwischen den Eigenschaften eines Songs, die sich auf die Benutzervorlieben und das entsprechende Audiosignal auswirken. Wir zeigen auch das jüngste Fortschritte im Tiefen Lernen sich gut in das Musikempfehlungsproblem transferieren lässt und man mit Hilfe von Tiefen CNN den traditionellen Ansatz deutlich übertrifft.

1 Introdution:

In den letzten Jahren hat sich die Musik Industrie immer mehr in die Richtung Digitalisierung, durch online Musikgeschäften und Streaming Service wie iTunes, Spotify, Grooveshark und Google Play gewandt. Als Resultat daraus wurde das Musikempfehlungsproblem immer interessanter, welches dem Zuhörer erlaubt neue Musik nach seinem Geschmack kennen zu lernen, weiter ermöglicht es Online Musikgeschäften ihre Waren dem richtigen Kunden anzubieten.

Obwohl das Empfehlungssystem schon sehr stark erforscht wurde, ist das Problem mit Musikvorschlägen die große Anzahl an verschiedenen Stilen und Genres, so wie die sozialen und geographischen Faktoren, die die Vorzüge eines Hörers beeinflussen. Die Zahl an Items die Empfohlen werden können ist gigantisch, vor allem wenn es sich um die Empfehlung individueller Lieder handelt. Diese Zahl kann dadurch reduziert werden, wenn man ein bestimmtes Album oder einen Künstler stattdessen vorschlägt. Allerdings ist das nicht immer mit dem benutzten System kompatible und es beachtet nicht den Fakt, dass die Musik eines Künstlers selten homogen ist, ein Zuhörer kann bestimmte Lieder lieber hören als andere.

Viele Empfehlungssysteme sind von Nutzungsmustern abhängig, die Kombination aus Musik die ein Benutzer konsumiert oder bewertet, gibt Information über die Vorlieben eines Nutzers und wie die Lieder untereinander zusammenhängen. Das ist die sogenannte kollaborativer Filterungsansatz. Ein anderer Ansatz ist die Nutzervorlieben von Item Inhalt und Metadaten vorherzusagen.

Der Konsens daraus ist, dass ein kollaborativer Filterungsansatz eine Inhalts basierte Empfehlung in der Regel überbietet. Dies ist aber nur möglich, wenn Daten zur Benutzung vorhanden sind. Der Kollaborativer Filterungsansatz hat das „Cold Start“ Problem, neue Inhalte die noch nicht angehört wurden, können nicht empfohlen werden. Zusätzlich dazu sind Inhalte, die nur für eine Nische an Zuhörern interessant ist, sind schwerer zu empfehlen, da es nur wenige Nutzungsdaten gibt. In vielen Gebieten und besonders in der Musik wird eine Mehrheit an verfügbaren Elementen gebildet, da das Konsumverhalten der Zuhörer einem Potenzgesetz folgt. Die Inhaltsbasierte Empfehlung ist davon nicht betroffen.

* 1. Content-based music recommendation

Musik kann anhand seiner Metadaten empfohlen werden. Informationen über den Künstler, Album und das Erscheinungsjahr sind normalerweise bekannt. Unglücklicherweise führt dies zu einer vorhersagbaren Empfehlung. Es ist nicht sehr sinnvoll Lieder eines Künstlers vorzuschlagen, den der Nutzer bereits kennt.

Man kann auch versuchen Musik zu empfehlen die wahrnehmbar ähnlich zum zuletzt gehörten Lied klingt, indem man die Ähnlichkeiten der Audiosignale misst. Dieser Ansatz benötigt eine Definition für geeignete Ähnlichkeitskennzeichen. Solche Kennzeichen sind oft aus dem Stegreif definiert, sie basieren aus bekannten Wissen über die Audiomusik, wodurch sie nicht mehr optimal für die Aufgabe der Musikempfehlung sind. Aus diesem Grund haben einige Forscher Nutzerpräferenzdaten verwendete um Ähnlichkeiten abzustimmen.

* 1. Collaborative filtering

Kollaborative Filter Methoden können Nachbarn- oder Modelbasiert sein. Die ehemaligen Methoden basieren auf Ähnlichkeitsmerkmalen zwischen Hörer und Element. Sie empfehlen Elemente, die andere Nutzer mit ähnlichen Vorlieben gehört haben, oder Titel, die Ähnlichkeiten mit bereits gehörten Elementen aufweisen. Modelbasierte Methoden versuchen ein Model über die verborgenen Eigenschaften über Nutzer und Elemente zu erstellen, welches als Vektor über die Eigenvektoren dargestellt wird. Eigenvektoren Modelle sind seitdem ihre Effizienz im Netflix Preis gezeigt wurde sehr populär.

* 1. The semantic gap in music

Eigenvektoren formen eine kompakte Beschreibung von unterschiedlichen Facetten des Geschmacks eines Nutzers und den entsprechenden Eigenschaften der Elemente. Um dies zu zeigen haben wir Eigenvektoren eines kleinen Datensatzes generiert und listeten Künstler, deren Lieder sehr positive und sehr negative Werte für jeden Faktor haben, in einer Tabelle. Diese Darstellung ist durchaus vielseitig und kann in verschiedenen Bereichen neben der Empfehlung benutzt werden. Da Nutzugsdaten für viele Lieder selten sind, ist es oft unmöglich die Faktorvektoren zuverlässig zu schätzen. Deshalb wäre es nützlich, wenn man dieses aus dem Musikaudio Inhalt vorhersage könnte.

Es gibt eine große semantische Lücke zwischen den Charakteristiken eines Lieds, die die Nutzerpräferenzen beeinflussen und des dazu gehörigen Audiosignals. Um Qualitative Eigenschaften wie Genre, Stimmung, Instrumentierung und die lyrischen Themen von einem Audio Signal extrahieren zu können, benötigt man ein starkes Model, das die komplexe Hierarchie Struktur von Musik erfasst. Zusätzlich ist es unmöglich bestimmte Eigenschaften, wie die Popularität eines Künstlers, deren Ruf und Lage, allein aus dem Audio Signal herauszubekommen.

Forscher des Bereiches Musik Information Retrieval beschäftigen sich mit dem Extrahieren dieser High-Level Eigenschaften der Musik. Sie vertrauen auf ein bestimmtes Set von Audiofeatures, wie die Mel-Frequenz-Cepstal-Koeffizienten, welche als Eingaben für eine einfache Klassifikator oder Regressor, wie SVMs oder lineare Regression, benutzt werden. Dieser traditionelle Ansatz wurde von einigen Autoren, die tiefen neuronale Netzwerke auf das MIR-Problem angewendet haben, in Frage gestellt.

In diesem Paper bemühen wir uns den Semantischen Unterschied von Musik zu überbrücken, idem wir ein tiefes Neurologe Netzwerk trainieren, Eigenvektoren von Musik Audio vorherzusagen. Wir bewerten unseren Ansatz durch einen industriellen Datensatz mit Audio-Auszügen aus über 380000 Liedern, und vergleichen diese mit dem Konventionellen Ansatz, dem „Bag-Of-Words“ Feature für jedes Lied. Wir bewerten inwieweit es möglich ist Charakteristiken, die die Benutzereinstellungen beeinflussen, direkt aus den Audiosignalen zu extrahieren, und bewerten die Vorhersage von unseren Modellen durch ein Musikempfehlungs Einstellung.

2. The dataset

Der eine Millionen Lieder Datensatz (MSD) ist eine Sammlung von Metadaten und vorberechneter Audiofunktionen für eine Million zeitgenössischer Lieder. Es gibt viele andere verfügbare Datensätze, die mit der MSD verknüpft sind, diese enthalten Texte, Cover-Lieder, Tags und Daten über das hörverhalten von Hörern. Dies macht den Datensatz für eine breite Palette an verschiedenen Musikinformationsabrufaufgaben nutzbar. Für unser Experiment sind zwei verlinkte Datensätze

Traditionelle Forschung im Bereich Musik Information Abruf (MIR), von größeren Datensätzen, war auf die Industrie limitiert, da es aufgrund von Lizenzproblem nicht möglich war ein große Sammlung Musik zu veröffentlichen. Die Autoren des MSD haben dieses Problem umgangen, indem sie vorab berechnete Funktionen anstatt der original Audiodateien zur Verfügung stellen. Unglücklicherweise sind die Audiofunktionen, die die MSD anbietet nur begrenzt nutzbar, und der Prozess mit dem sie erhalten wurden ist schlecht dokumentiert. Der Funktionsumfang wurde zwar durch Rauber erweitert, allerdings ist das Fehlen roher Audiodaten oder zumindest eine Darstellung auf mittlerer Ebene, immer noch ein Problem. Trotz allem konnten wir 29 Sekunden Audio-Clips vom Datensatz auf 7digital.com bekommen.

Aufgrund der Größe, ist es möglich am der Forschung am Musikempfehlungsproblem realistischer Einstellungen zu bekommen. Es ist auch erwähnenswert, dass das Geschmacksprofil ist einer der größten kollaborativen Filter-Datensätze, die heute öffentlich verfügbar sind.

3. Weighted matrix factorization

Das Geschmacksprofil enthält die Anzahl der Wiedergaben pro Song und Benutzer, welches eine Form von impliziten Feedback ist. Wir wissen zwar wie oft ein Benutzer ein Lied im Datensatz angehört hat, aber er diesen nicht explizit bewertet. Allerdings kann man davon ausgehen, dass ein Nutzer vermutlich öfters dasselbe Lied anhört, wenn ihm dieses gefällt. Falls ein Benutzer nie ein Lied angehört hat, kann dies an mehreren Faktoren liegen: zum Beispiel kennen sie das Lied nicht, oder sie mögen das Lied nicht. Diese Einstellung ist nicht kompatibel mit traditioneller Matrix Faktorisierung, die auf die Vorhersagen von Bewertung zielen.

Wir benutzen den Gewichtete Matrix Faktorisierungsalgorithmus (WMF), dargestellt durch Hu, um eine Eigenvektor Repräsentierung aller Nutzer und Elementen im Geschmacksprofil. Es handelt sich um eine modifizierten Matrix Faktorisierungsalgorithmus, der auf implizite Feedback-Datensätze gerichtet ist. Es sei *rui*der Zähler für einen Benutzer *a* und Lied *i*. Für jedes Nutzer-Elementen Paar definieren wir eine Präferenzvariable *pui* und eine Vertrauensvariable *cui* (I(x) ist die Indikator Funktion, alpha und epsilon sind Hyperparameter).

*pui* = I(*rui* > 0);

*cui* = 1 + alpha log(1+ epsilon-1 *rui*)

Die Präferenzvariable zeigt ob ein Benutzer u jemals ein Lied i angehört hat. Wenn es 1 ist, gehen wir davon aus, dass der Nutzer das Lied mag. Die Vertrauensvariable misst wie sicher wir uns über diese Aussage sind. Es ist eine Funktion des Zählers, da Lieder mit einem höheren Zähler eher bevorzugt werden. Wenn ein Lied nie angehört wurde, hat die Vertrauensvariable einen niedrigen Wert, da dies der Fall mit den wenigsten Informationen ist.

Die WMF Objektfunktion ist gegeben durch:



Gamma ist ein Regulierungsparameter, *xu* ist der Eigenvektor der Benutzer *u* und *yi* ist der Eigenvektor der Lieder i. Es besteht aus einem Vertrauensgewicht des mittleren quadratischen Fehlers und dem Regulierungsterm L2. Zu beachten ist, dass sich die erste Summe über alle Nutzer und Lieder erstreckt: im Gegensatz zu Matrix Faktorisierung für eine Rating-Vorhersage, in der Benutzer-Objekt Kombinationen ohne Bewertung verworfen werden, wodurch alle möglichen Kombinationen berücksichtigt werden. Da die Nutzung des Stochastischen Gradientenabstieg zu Optimierung des Datensatzes nicht praktikabel ist, haben wir den von Hu vorgeschlagenen Optimierungsmethode (ALS) benutzt.

4. Predicting latent factors from music audio

Das Vorhersagen eines Eigenvektors eines gegebenen Liedes, des dazugehörigen Audio Signals ist ein regressives Problem. Es erfordert das Erlernen einer Funktion, die eine Zeitreihe einen Vektor reeller Zahlen zuordnet. Wir bewerten dazu zwei Methoden um dies zu erreichen: in der Ersten folgt man dem traditionellen Ansatz in MIR, indem man lokale Features eines Audiosignals in einer „Bag-of-Words“ Darstellung verbindet. Es kann eine beliebige traditionelle Regressionstechnik verwendet werden, um dieser Merkmaldarstellung die Faktoren zuzuordnen. Die andere Methode besteht in der Verwendung eines tiefen neuronalen Netzwerks.

Eigenvektoren, die durch die Anwendung von WMF auf die verfügbaren Nutzdaten erhalten werden, werden als Grundwahrheit verwendet um das Vorhersagemodell zu trainieren. Es sollte beachtet werden, dass dieser Ansatz mit jedem Typ eines Eigenvektor-Modell kompatibel ist, dass für große implizite Rückkopplungsdatensätze geeignet ist. Wir haben uns für WMF entschieden, weil es dafür ein effizienter Optimierungsprozess existier.

4.1 Bag of words representation

Viele MIR Systeme brauchen eine der folgenden Funktionsextraktionskanäle, um Musikaudiosignale in eine Darstellung mit fester Größe zu konvertieren, welche als Eingabe für einen Klassifikator oder Regressor verwendet werden kann.

* Extrahieren von MFCCs aus den Audiosignalen: Wir berechnen 13 MFCCs aus Fenstern von 1024 Audioausschnitten, welches 23 ms bei einer Abtastrate von 22050 Hz entspricht und einer Sprunggröße von 512 Proben. Wir berechneten auch die Differenz der ersten und zweiten Ordnung, welche insgesamt 39 Koeffizienten ergab.
* Vektorquantisierung der MFCCs: Wir haben ein Wörterbuch aus 4000 Elementen mit dem K-Durschnittsalgorithmus erlernt und allen MFCC-Vektoren den nächstliegen Mittelwert zugeordnet.
* Vereinigt sie in einer „Bag-of-Words“-Darstellung: Für jedes Lieder haben wir gezählt, wie oft der Mittelwert ausgewählt wurde. Der daraus erzeugte Zählvekor ist eine Merkmalsdarstellung des Liedes.

Dann haben wir die Größe der Darstellung mit Hilfe von PCA reduziert (wir haben genug Komponenten um 95% der Varianz abzudecken) und verwendeten die lineare Regression und ein mehrschichtiges Perzeptron mit 1000 versteckten Einheiten darüber um Eigenvektoren vorhersagen zu können. Wir haben es auch als Eingabe für den MLR Algorithmus verwendet, um eine Ähnlichkeitsmetrik für inhaltsbasierte Empfehlungen zu erlernen. Diese Erkenntnisse wurden als Basis für unsere Musikempfehlungsexperimente, die in 5.2 beschrieben werden, genutzt.

4.2 Convolutional neural networks

CNNs wurden erst vor kurzen verwendet, um den Stand der Technik in Spracherkennung und Bilderkennung zu verbessern. Dabei scheinen drei Bestandteile im zentralen Focus des Erfolgs dieses Ansatzes zu stehen:

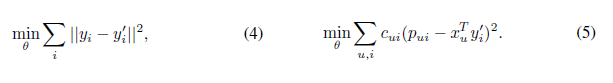
* Das Benutzten von RELUs anstatt nichtlineare Sigmuid-Funktionen führt zu einer schnelleren Konvergenz und reduziert das Verschwinden des Gradienten-Problems, welches ein Problem traditionaler vielschichtiger Netzwerken ist.
* Parallelisierung wir genutzt um schneller Trainieren zu können, wodurch große Modelle in angemessener Zeit trainiert werden können. Wir haben die Theano Bibliothek für GPU-Beschleunigung genutzt.
* Es werden viele Trainingsdaten benötigt, damit man große Modelle mit vielen Parametern nutzen kann. Die MSD enthält genug Daten um ein großes Modell effizient zu trainieren.

Wir haben auch die Verwendung der Dropout-Regulierung untersucht, aber dies brachte keine signifikanten Verbesserungsergebnisse.

Wir haben zunächst eine Zwischenzeitliche Frequenzdarstellung des Audiosignals extrahiert, um dieses als Eingang des Netzwerkes zu nutzen. Wir verwendeten log-komprimierte-Mel-Spektogramme mit 128 Komponenten und derselben Fenster und Sprunggröße die wir für die MFFCs benutzten (1024 bzw. 512 Audio-Frames). Das Netzwerk wurde auf 3 Sekunden Abschnitte von zufälligen Audioclips trainiert. Dies wurde primär genutzt um das Training zu beschleunigen. Um den Eigenvektor eines ganzen Liedes vorherzusagen, haben wir die Vorhersagen aufeinanderfolgender Ausschnitte gemittelt.

Konvolutionelle neuronale Netzwerke eignen sich besonders um Vorhersagen über Eigenvektoren von Musikaudio zu machen, da Zwischenmerkmale zwischen verschiedenen Faktoren geteilt werden und weil sie durch ihre hierarchische Struktur bestehend aus Merkmalsextrationsebenen und Pooling-Layern auf mehreren Zeitachsen gleichzeitig arbeiten können.

4.3 Objective functions

 Eigenvektoren sind reell wertig, wodurch das einfachste Ziel die Minimierung der Vorhersage durch den MSE. Alternativ dazu können wir auch das Gewicht der Fehlervorhersage (WPE), der WMF-Zielfunktion, minimieren. Sei *yi* der Eigenvektor für ein Lied i, enthalten in WMF, und *y0i* die entsprechende Vorhersage des Models. Die Zielfunktionen:

5. Experiments

5.1 Versatility of the latent factor representation

Um die Vielseitigkeit von Eigenvektoren zu demonstrieren, vergleichen wir sie mit Audio-Features in einer Tag-Vorhersage. Tags beschreiben eine große breite an verschiedenen Aspekten eines Lieds, wie das Genre, Instrumentierung, Geschwindigkeit, Stimmung und das Erscheinungsjahr.

Wir haben WMF benutzt um einen 50 Dimensionalen Eigenvektor für alle 9330 Lieder in der Untergruppe zu erhalten und trainierten ein logisches Regressionsmodell, um die 50 beliebtesten Tags aus dem Last.fm Datensatz für jedes Lied vorherzusagen. Zudem haben wir dein logisches Regressionsmodell der „Bag-of-Words“-Darstellung über die Audiosignale trainiert, welche zunächst in seiner Größe durch PCA reduziert wurde. Wir verwenden eine 10-fache Kreuzvalidierung und berechnen die durchschnittliche Fläche mit Hilfe der ROC-Kurve (AUC) über alle Tags. Dies führte zu einer durchschnittlichen AUC von 0.69365 für die Audio basierte Vorhersage und 0.86703 für die Eigenvektor Vorhersage.