4J1-1

機械学習を用いた楽曲に対する感性推定の手法

Estimating Affects on Music with Machine Learning

大槻良祐 *1 Ryosuke Otsuki　福井健一 *2 Ken-ichi Fukui　森山甲一 *2 Koichi Moriyama　大谷紀子 *3 Noriko Otani　沼尾正行 *2 Masayuki Numao

*1大阪大学大学院情報科学研究科情報数理学専攻
Department of Information and Physical Sciences, Graduate School of Informartion Science and Technology, Osaka University

*2大阪大学産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

*3東京都市大学メディア情報学部
Faculty of Informatics, Tokyo City University

These days, we are in information overload, so recommendation system for each individual are required. We have focused on the system that recommend music adapting to the user's affects. In this paper, we analysis Moodswings dataset which represent subjects' affects when they listen music. And then, we formulate estimating human affect with machine learning. This lead to the music recommendation system more accuracy.

1. はじめに

近年，ビッグデータ等のキーワードが話題になるなど情報過多の状況が加速しており，個人に適した情報を提供する事が求められている．その中でも，感性を考慮することでコンテンツの推薦精度を上げようという研究が盛んに行われている.[1] 本研究室では楽曲が人間の感性に与える影響に着目し，感性に基づいた自動作曲システム [2] や，楽曲聴取時の脳波データによる感性推定器の構築 [3] 等，様々な研究を進めてきた．

本研究では楽曲推薦システムに用いる感性モデルに対して新たなアプローチを提案する. 従来の楽曲聴取時の感性モデルに関する研究では，楽曲に対する感性を数段階で評価する SD 法 (semantic differential method) 等が用いられていたため連続的な評価が困難であった．Schmidt ら [7] の研究では Russell の AV 空間 [4] を用いることで感性に対する連続的な評価を可能にし，より複雑な感性モデルを獲得している．この感性モデルはある楽曲に対して被験者全体の感性がどのように変化するかを推定するモデルであり，例えば ある楽曲聴取中に被験者全体の傾向として“楽しい”→“落ち着く”のような遷移が起きるだろうといった推定を行う．しかし，Schmidt らの研究でも示されている通り，一般的に同じ楽曲を聴取しても異なる感性が想起されるので，個人に適した楽曲推薦を行う場合は，この感性モデルを直接使用することは難しいと考えられる．最も単純な解決策として, 被験者毎に感性モデルを獲得することが考えられるが，Shmidt らの手法では被験者への負担が大きく，また同じ被験者が同じ楽曲を聴取しても日時により結果が異なるため，推薦システムとしてあまり実用的でなくなってしまう．そこで，本研究では被験者全体のデータを分類することで，同一楽曲から複数の結果を推定するように Schmidt らのモデルを拡張する．これは，教師ありの時系列データを特徴的な遷移パターン毎に分類し推定を行う問題に相当する. 例えばある楽曲聴取時には被験者により“楽しい”→“落ち着く”，“悲しい”→“優しい”等の遷移パターンが起こることを推定する. 楽曲推薦システムでは被験者からのフィードバックにより都度被験者の感性がどの遷移パターンに近しいかを推定し推薦を行う．

本研究の結果としては，適切なクラスタ数の決定を可能にし，またいくつかの特徴的な感性モデルの重み (楽曲聴取時の AV 値を決定し，感性に相当する) を抽出することに成功した．

連絡先: 大槻良祐 大阪大学 産業科学研究所沼尾研究室
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1
Tel:06-6879-8426 Fax:06-6879-8428
E-mail:otsuki@ai.sanken.osaka-u.ac.jp

2. 関連研究

楽曲と感性に関する研究では様々な手法が用いられているが，主に楽曲・感性の表現方法により大別される．感性の表現としては，アンケートを用いて数段階で表現する SD 法 (semantic differential method) や，脳波・心拍等の生理信号がよく用いられている.[3] 楽曲の表現としてはコード・リズム・和音進行といった楽曲理論に関するものや，MFCC(mel frequency cepstral coefficients)[5] といった音声認識の分野でよく取り上げられる特徴量が用いられることが多い．

ここでは，本研究で用いた Russell の AV 空間と Schmidt らの研究について説明を行う．

2.1 Russell の AV 空間

本研究では感性の表現として Russell の AV 空間 [4] に 1 秒毎の座標値が与えられたデータセットを用いている．Russell の AV 空間は 図 1 のように energetic-silent(Arousal) positive-negative(Valence) の二次元平面で感性を表現するものであり，それぞれの領域ごとにある感性に対応していると考えられている．Russell の AV 空間を用いると時間を追った感性の推移を得ることが出来る．本研究で用いたデータセットの一例を図 1 に示す．また，Russell の AV 空間を用いた Feeltrace や Moodswing[6] といった感性情報取得のためのソフトウェアによってインターネット上で世界中から感性データを集めることも可能となっている．

2.2 感性モデル

Russell の AV 空間による楽曲聴取時の感性モデルは，Schmidt ら [7][8] により非常によく研究されている。Schmidt らの研究では MFCC 等の特徴量 (及びそれらの組み合わせ) を入力として，最小二乗回帰（Least-Squares Regression）やカルマンフィルターを用いて感性モデルを獲得している．

最小二乗回帰では $\boldsymbol{x}$ を入力，y_i を出力 (Arousal-Valence 値) とした以下の様な線形回帰モデルで感性モデルが表現される．

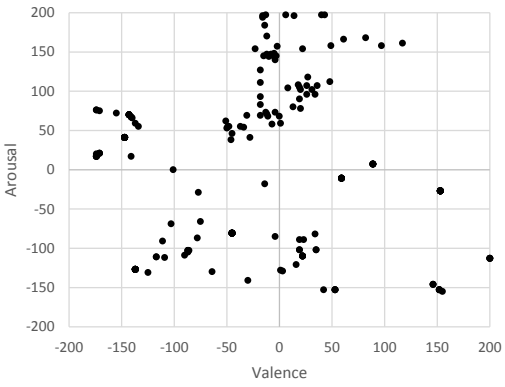

図 1: 楽曲の AV 値分布

$$y_i = \boldsymbol{x}_i \boldsymbol{w} + \mathcal{N}(0,\sigma^2) \tag{1}$$

ここで $\boldsymbol{w}$ は重み，$\mathcal{N}(0,\sigma^2)$ は正規分布を表す．データセットから重みを学習することにより，各秒数での Arousal-Valence 値を推定可能としている．このモデルでは重み $\boldsymbol{w}$ が決定されれば全ての楽曲の Arousal-Valence 値を推定できるので，重み $\boldsymbol{w}$ が楽曲に対する感性に相当すると考えられる．

3. 提案手法

Schmidt らの研究により楽曲から感性を定量的に推定する感性モデルを獲得できたが，Schmidt らの手法ではある入力に対して単一の値しか出力されない．よって時系列データである楽曲からは一つの遷移パターン (Arousal-Valence 空間における) が得られる．一般的に同じ楽曲を聴取しても被験者ごとに遷移パターンは異なるが，被験者毎に Arousal-Valence 値を取得することは負担が大きく，また同じ被験者が同じ楽曲を聴取しても日時により結果が異なるため，被験者ごとに感性モデルを獲得することは難しいと考えられる．そこで本研究ではデータセットを N クラスタに分類することで，楽曲に対し N 通りの遷移パターンを出力する (入力に対し N 個の値を出力する) 感性モデルを提案する．これは，Schmidt らのモデルにおいて，分割された各データセット毎に重み $\boldsymbol{w}$ を学習することに相当する．

3.1 仮定

本研究では推薦システムに感性モデルを適用することを考えており，初めに以下の仮定を行った．

- 楽曲に対する感性は N 通りで表現することが出来，ユーザーは常にどれかに所属している (これは N 個の重み $\boldsymbol{w}$ で全てのユーザーの遷移パターンを説明可能なことを仮定している)
- 少なくとも楽曲を聴取している間は，ユーザーが属するクラスタは変更しない

一つ目の仮定は，推薦システムをクラスタモデルとして扱うための仮定であり，二つ目の仮定は，システムを実用的に運用するための仮定である。一つ目の仮定により，推薦システムが適切な楽曲推薦を行うためには，ユーザーの属するクラスタを特定するような機能が必要になる．この時，各秒数毎にクラスタの推定を行うと推定数が指数的に発散してしまうため，二つ目の仮定を行った．

3.2 手法

データセットを N クラスタに分割し，それぞれで学習を行うことで楽曲ごとに N 通りの遷移パターンを出力する感性モデルを獲得できるが，データの分割方法により結果は大きく変わる．ここで，最も良いデータの分割方法は学習後に各モデルでの誤差が最小となる手法であると考えられる．本研究では，階層型クラスタリングを用いることで各モデルの誤差を減少させており, 以下にその手法を示す．

まず，モデルとしては Schmidt らの手法より表現力の高い多層パーセプトロンを用いる．多層パーセプトロンは入力層・隠れ層・出力層を持つ 2 層のニューラルネットワークであり，以下の誤差関数を誤差逆伝播法により最小化することで学習を行う.(ここで，$\boldsymbol{t}$ は教師データ，$\boldsymbol{y}$ はフィードフォワードネットワーク関数を表す)

$$error(\boldsymbol{w}_i) = \sum_i (||\boldsymbol{t}_i - \boldsymbol{y}(\boldsymbol{x}_i, \boldsymbol{w}_i)||^2) \tag{2}$$

この時，活性化関数にシグモイド関数を適用することで関数 $\boldsymbol{y}$ が非線形関数となり，線形回帰モデルより表現力が高くなることが知られている．多層パーセプトロンを適用した階層型クラスタリングの手順を以下に示す．まず，二つ目の仮定により楽曲聴取時の各時系列データが同じクラスタとなる．本研究で用いたデータセットは 15 秒間毎なので各初期クラスタは 15 秒の時系列データとなる．

1. 初期クラスタ毎に多層パーセプトロンで学習を行い，それぞれのクラスタ u_i の重み $\boldsymbol{w}_i$ を得る．
2. 得られた重みをもとに各クラスタ u_i, u_j 間の非類似度 d_{ij} を以下の計算式より求める．

$$error_{ij} = \frac{1}{n_{u_j}} \sum_k (||\boldsymbol{t}_{jk} - \boldsymbol{y}(\boldsymbol{x}_{jk}, \boldsymbol{w}_i)||^2) \tag{3}$$

$$d_{ij} = d_{ji} = min(error_{ij}, error_{ji}) \tag{4}$$

ここで，$error_{ij}$ はクラスタ u_j のデータをクラスタ u_i に所属させた時の誤差を表す.(n_{u_j} はクラスタ u_j のデータ数)

3. 全ての非類似度の中で最小のものを取り出し，それが設定した閾値よりも低い場合，二つのクラスタを統合する．この時，新しいクラスタの重みは $error_{ij} > error_{ji}$ であれば $\boldsymbol{w}_j$ を採用する.(クラスタ数が 1 になればクラスタリングを終了する.) 最小の非類似度が，閾値よりも大きい場合には各クラスタ毎に多層パーセプトロンの学習を行う．
4. 2-3 を繰り返す. 再学習後も最小の非類似度が閾値より大きい場合はクラスタリングを終了するか閾値を増加させる．

上記の手順により，設定された閾値以下の分散をもつ複数のニューラルネットが構築される．この時，手順 1 で 15 秒間毎の時系列データで初期クラスタを作成するため, 時系列的な変化にも対応できることが期待される．懸念点としては，手順 1 ではデータ数が少くなるため過学習になることが想定されるが，手順 2-3 でより汎化性能の高い重みが選択されるためある

程度抑えられると考えられる. ここで，理想的には非類似度はクラスタ u_i, u_j のデータで多層パーセプトロンの学習を行った結果の誤差関数の値を用いるべきだが，計算量を考慮してより簡易的な方法を採用した.

4. 検証方法

4.1 想定する楽曲推薦システム

検証を行う前に楽曲推薦システムの仕様をある程度決めておく必要がある. ここでは単純に被験者が楽曲聴取後に Arousal-Valence 値を入力するシステムを考える. この時 最も誤差が少ないクラスタに属していると推定すると考える.

4.2 データセット

データセットは Scmidt らの研究と同じものを用いた。データセットにはあらかじめ行われていた被験者実験により 1 秒毎の Arousal-Valence 値が与えられている.240 曲に対し各 15 秒間毎，それぞれ被験者が 10〜20 人程，約 60 000 点のデータセットとなっている. また, 初期クラスタ数は 4062 である.

4.3 推定器の設定

全ての推定器の学習において，出力は各秒数毎の Arousal-Valence 値を -1〜1 にスケールを変化したものを用いた. 入力は，混合密度ネットワークでは出力の前 3 秒間での MFCC を 1 秒ずつ平均し，各次元毎に正規化を行ったものを用いた.MFCC は 20 次元の値であるので，入力次元は 60 次元となる. また隠れ層は 30 次元とした. 提案手法では過学習を抑制するため，混合密度ネットワークの入力を Auto Encoder[9] で 30 次元に次元削減したものを用いた. 隠れ層数は 10 次元とした.

4.4 検証方法

分割数を 10 とした交差検定を用いて検証を行った. ただし，推薦システムが正しいクラスタを選択した際の精度を求めたいので，各 15 毎間の時系列データで最も誤差の少ない遷移パターンの結果を用いた. また，提案手法との比較のために，単一の楽曲から N 通りの遷移パターンを出力できる混合密度ネットワーク (mixture density network)[10] も用いた.

5. 結果

5.1 クラスタ数

提案手法では閾値を設定するが，これは分散と対応する. ここで，階層型クラスタリングを行った際の閾値とクラスタ数の結果を図 2 に示す. また，15 秒間の時系列データの数が 5 未満のクラスタを外れ値として除いた場合の値も載せている. 外れ値を除いた場合ではクラスタ数が初めは増加し ピークの後は減少していることが読み取れる.

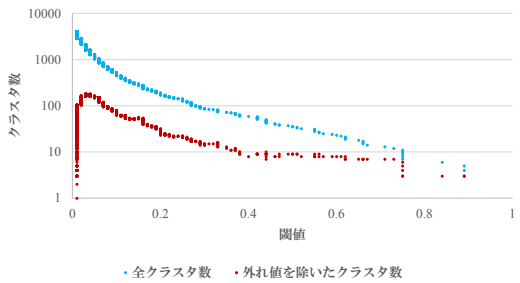

図 2: クラスタ数と閾値の関係

5.2 汎化性能

各手法についての汎化性能を図 3 に示す. ここでは推薦システムを考慮し，回帰問題としてクラスタ毎の各秒数での誤差 ($||\boldsymbol{t}_i - \boldsymbol{y}(\boldsymbol{x}_i, \boldsymbol{w}_i)||$) の平均を用いた.Schmidt の研究では各秒数毎の全ての被験者の平均との距離を評価基準としていたので，Scmidt らの論文での値とは少し異なっていることに注意する必要がある. また，混合密度ネットワークではクラスタ数の増加に従い計算量も増加するので，16 クラスタまでしか計算を行っていない. 値としては Arousal-Valence 値を -1〜1 にスケールを変化させているので 誤差が 0.2 の場合は全体の 10%程度の誤差となる.

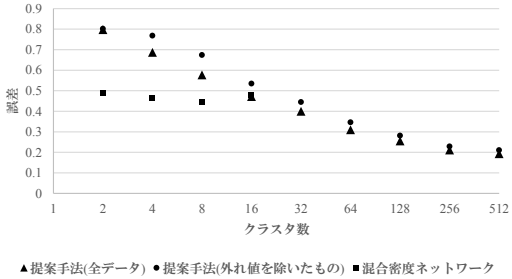

図 3: クラスタ数と汎化性能の関係

6. 考察

6.1 汎化性能

クラスタ数が低い場合は，混合密度ネットワークの精度が高いことがわかる. しかし，クラスタ数が増加するにつれ階層型クラスタリングは順当に精度を上げているのに対し，混合密度ネットワークではほとんど変化が無い. クラスタ数が少ない場合に差が出ている原因は外れ値を除いた場合にクラスタ数に差が出ていること，階層型クラスタリングでは混合密度ネットワークと違い局所的にクラスタを作っていることが考えられる. 混合密度ネットワークがクラスタ数が増加した場合に汎化性能が上がらない原因は，局所解に収束している可能性と，時系列的な流れを捉えられていないためと考えられる.

また，図 2 を見るとクラスタ数が 128 の場合では，外れ値を除いたクラスタ数は 21 であることがわかる.(交差検定では 21〜31 となった) この時，図 3 を見ると全データでのクラスタ数 32 の誤差は 0.400，クラスタ数 128 の外れ値を除いた誤差は 0.282 となっており後者の方が汎化性能が高いことがわかる. 他の値でも同様に，外れ値を除いたクラスタ数での汎化誤差がどれも倍のクラスタ数のものより改善されている. このことから，外れ値を除くことで特徴的な感性モデルの重み (楽曲聴取時の AV 値を決定し，感性に相当する) を抽出出来ると考えられる.

6.2 クラスタ数

結果よりクラスタ数と誤差の関係が明らかとなったので，システムの目的に応じて許容できる誤差を選択することで適切なクラスタ数を決定出来る. 例えば 15%以上の誤差を許容できなければ 128 クラスタの場合の外れ値を除いたクラスタを用いれば良い.

6.3 課題

本研究で用いた手法は階層型クラスタリングであるので 計算量が大きく より大きなデータセットへの適応が難しい. また

より適応的な推定を考えるのであれば オンラインの学習にも対応できる方が望ましいと考えられる.

7. おわりに

本研究では，感性を考慮した楽曲推薦システムの構築を目標とし，楽曲に対する感性のパターン毎に推定できるよう既存手法の拡張を試みた. 結果として，目的に応じたクラスタ数の決定・代表的な感性のパターンの抽出に成功した.

参考文献

[1] Tanaka, M., Hiroyasu, T., Miki, M., Ya-sunari, S., and Yoshimi, M.“Automatic Generation Method to Derive for the Design Variable Spaces for Interactive Genetic Algorithms” Proc. IEEE World Congress on Computational Intelligence, 2010.

[2] Masayuki Numao, Shoichi Takagi, and Keisuke Nakamura.“Constructive Adaptive User Interfaces - Composing Music Based on Human Feelings”, Proc. Eighteenth National Conference on Artificial Intelligence (AAAI-02), pp.193-198, 2002.

[3] Rafael Cabredo, Roberto Legaspi, Paul Salvador Inventado, and Masayuki Numao. “Discovering Emotion-Inducing Music Features Using EEG Signals”, Journal of Advanced Computational Intelligence and Intelligent Informatics, 17 (3). pp.362-370, 2013.

[4] J. A. Russell,“A Circumplex Model of Affect,” Proc. J. Personality Social Psychology, vol.39, pp.1161-1178, 1980.

[5] B. Logan,“Mel frequency cepstral coefficients for music modeling,” Proc. the International Symposium on Music Information Retrieval (ISMIR), 2000.

[6] Jacquelin A. Speck Erik M. Schmidt Brandon G. Morton and Youngmoo E. Kim “A Comparative Study of Collaborative VS. Traditional Musical Mood Annotation” Proc. 12th International Society for Music Information Retrieval Conference pp.549-554 2011.

[7] Schmidt E. M. and Kim Y. E.“Prediction of Time-Varying Musical Mood Distributions from Audio.” Proc. 10th International Society for Music Information Retrieval Conference 2010.

[8] Schmidt E. M. and Kim Y. E.“Modeling and Predicting Emotion in Music.” Proc. Music Mind and Invention Workshop, 2012.

[9] G.E. Hinton and R.R. Salakhutdinov,“Reducing the Dimensionality of Data with Neural Networks,” Proc. Science, vol.313. no.5786, pp.504-507, 2006.

[10] Bishop,“Neural Networks for Pattern Recognition,” Proc. Oxford University Press, 1995.