The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

非負値多重行列因子分解の因子行列を用いたクラスタリングと決定木学習によるオフィスの入退室データの分析

An Analysis of Entry and Exit Data in Office by Decision Tree Learning Using Clustering Factor Matrix from Non-negative Multiple Matrix Factorization

小島世大 Seidai Kojima　石榑隼人 Hayato Ishigure　坂田美和 Miwa Sakata　武藤敦子 Atsuko Mutoh　森山甲一 Koichi Moriyama　犬塚信博 Nobuhiro Inuzuka

名古屋工業大学 大学院 工学研究科 情報工学専攻
Department of Computer Science, Graduate School of Engineering, Nagoya Institute of Technology

Recently, IC card systems are popular and their log data are used for analyzing human behaviors. In this paper, we extract user behavior patterns using Non-negative Multiple Matrix Factorization (NMMF) and propose an analysis method to analyze patterns and attribute information by decision tree learning using clustering factor matrix. We examine our proposed method using actual entry and exit data and confirm the effect.

1. はじめに

近年 IC カードの普及に伴い、IC カード利用履歴を用いた人の行動分析の研究が増えている [1][2]。さらに、電子錠と IC カードにより、人の入退を制御する入退室管理システムの導入が増えている。入退室管理システムは、電子錠に備えられたカードリーダにかざされた IC カードの情報を読み取り通行履歴として記録する。現在では、企業は単純な入退室制御だけではなく、社員や部門毎の入退室傾向の分析や社員がどの部屋にいるのかを把握したいなどの要望がある [3]。また、近年、センサノード及びネットワークの技術が急速に発展している。環境に設置されたセンサノードと人が身に着けるウェアラブルセンサーの併用は、企業におけるオフィス内のワーカーの行動のモニタリングへ利用されている。センサデータを用いたオフィスワーカーの行動分析 [4] もされているが、これには特別な装置を必要とする。これまでに我々はクラスタリングによる入退室データからの移動時間パターンの分析をしてきた [5][6]。しかし、移動時間パターンの中で解釈が困難なものが存在した。そこで本研究では、非負値多重行列因子分解（Non-negative Multiple Matrix Factorization,NMMF)[7] を用いてユーザの行動パターンを抽出し、行動パターンと属性情報の関係を分析する手法を提案する。ＮＭＭＦを用いることで移動時間の特徴量だけではなく、部屋間の組に関する特徴量の行列を加えた社員の行動パターンの分析が可能となる。最後に提案手法を用いて多くの組織で普及が進んでいる入退室管理システムから得られる入退室データの分析を行う。

2. 関連研究

IC カードの利用者数の増加により、IC カード利用履歴データが人の行動分析手段として注目を集めている。 鈴木ら [2] は交通 IC カードの利用履歴を用いて人が駅の改札口を出て、次の改札口に入るまでの間にその人の滞在目的があるとし、生活パターンを定量的に表す生活行動属性を提案した。 カード利用者の生活圏 (駅) によらず似た生活パターンを持つ人を抽出しマーケティング等へ活用し得ると示した。また、嶋本ら [1] は英国・ロンドンで導入されている Oyster Card の 4 週間分の利用履歴データを用いて公共交通の変動を把握した。料金支払い形態に応じて利用者の利用属性を分類することで，1 人あたりの利用回数の変動の 4 割以上を説明できることを示した。

幸島ら [8] は消費者行動パターン抽出の為に、従来の NMMF にユーザや商品などの属性情報に関する入力行列を加えた、属性情報を考慮した非負値多重行列因子分解法（Non-negative Micro Macro Mixed Matrix Factorization,NM4F）の提案をした。提案手法と従来の NMMF の定量評価や因子行列をクラスタに分けて提案手法の定性評価を行っていた。

古川ら [9] はインターネットの YouTube にアップロードされた国内外のチェリスト 46 名に対して、著者が目視により様々な属性についてのデータを獲得し、クラスタリングと決定木分析などのデータ分析の手法を用いてチェリストの分類を試みた。

本研究では NMMF により抽出されたパターンと属性情報の関係を分析できる方法を提案する。本提案はＮＭＭＦにより得られた因子行列をクラスタに分けて属性の分析をする点で [8] と似ているが、クラスタリング結果を決定木学習に用いる点で異なる。また、クラスタリングと決定木分析を用いる点で [9] と似ているが、クラスタリングの特徴量ベクトルに因子行列を用いる点で異なる。

3. 提案手法

NMMF の因子分解結果を用いたクラスタリングと決定木学習によるユーザの頻出パターンと属性情報の関係を分析する手法を提案する。提案手法の流れを図 1 に示す。ＮＭＭＦとは複数の入力行列を同時に分解し、頻出パターンを同時に抽出する方法である。本提案の想定する入力行列は購買履歴データや入退室データなどから得られるユーザの購買パターンや行動パターンなどを表現する複数の特徴行列である。購買データであればユーザ×商品の購買回数の行列とユーザ×場所の訪問回数のデータである。まず NMMF により複数の入力行列から頻出パターンを抽出する。ユーザの属性情報との関係を分析する為に、因子分解により得られた因子行列の各行を特徴量ベクトルとし、クラスタリング手法を用いてユーザのグループ分けを行う。クラスタリングにより得られたグループ分け結果を元に、説明変数を属性情報とし、目的変数をクラスタ（グループ）に属するユーザを 1、その他のユーザを 0 とし、各クラスタ（グループ）に対して決定木学習を用いて分類木を作成する。各クラスタ中心をクラスタ（グループ）に含まれるユーザの特徴とし、因子行列と分類木を照らし合わせることで、頻出パターンと属性情報の関係を分析する。

連絡先: 小島世大, 名古屋工業大学 大学院, 〒 466-8555 愛知県名古屋市昭和区御器所町, s.kojima.571@nitech.jp

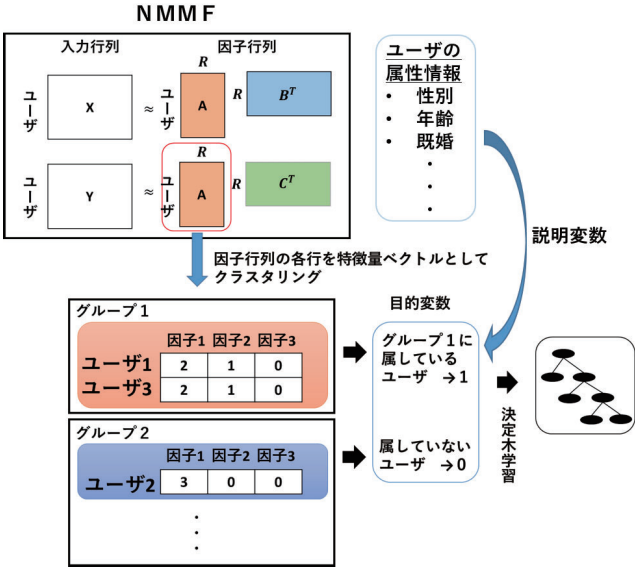

図 1: 提案手法の流れ

4. 提案手法を用いた入退室データの分析

4.1 入退室データ

本研究では (打刻日、打刻時刻、場所、操作、社員 ID) の 5 つの属性で構成された入退室データを扱う。操作には、部屋を退室もしくは入室したかが記録されている。移動時間に着目するために、退室と入室がセットになっているものを結合し、(移動開始日、移動開始時刻、移動時間、移動元、移動先、社員 ID) の 6 つの属性で構成されたデータ構造（以後移動データと呼ぶ）に変換する。移動時間は（入室時刻 – 退室時刻）で計算する。分析には協力企業の 2016 年 6 月の 332 名の移動データの内、移動時間が 90 分以内の移動のみを用いた。

4.2 提案手法に用いる入力行列の算出

社員ごとの移動時間に関する特徴量を表す行列（社員×移動時間）と場所に着目した部屋間の組に関する特徴量を表す行列（社員×部屋間の組）の二つの特徴行列を算出をする。

4.2.1 移動時間特徴行列 X

移動時間に関するの特徴量を表す行列 X（社員×時間区分）を算出する。社員 i の移動時間特徴行列の要素 X_{ij} は時間区分 j（時間帯及び移動時間の区分）の移動回数の割合とする [5]。分析に用いる時間区分を 9 種類 (時間帯 (午前・昼休憩・午後の 3 次元) ×移動時間 (0-5 分・5-20 分・20-90 分の 3 次元)) で表した（表 1 左）。

4.2.2 部屋間の組特徴行列 Y

部屋間の組に関する移動の特徴行列 Y（社員×部屋間の組）を算出する。社員 i の部屋間の組特徴行列の要素 Y_{ik} は部屋間の組（ex, A 室から B 室 →(A,B)）の移動回数の割合とする。例えば、ある社員が A 室から B 室の移動を 10 回しており、ある期間の総移動回数が 10 回の場合、部屋間の組特徴行列の要素の値は 1.0 となる。入力行列 Y の部屋間の組は 6 種類（a: 移動前と移動後同じ部屋、b:移動前と移動後同じビル、同じ階で違う部屋、c:移動前と移動後同じビルで違う階、d:第 1 ビルと第 2 ビルの間の移動、e:第 1 ビルと第 3 ビルの間の移動、f: 第 2 ビルと第 3 ビルの間の移動）で表した（表 1 右）。

表 1: 時間区分と部屋間の組

時間区分			部屋間の組	
記号	時間帯	移動時間	記号	
11	午前	0–5 分	a	移動前と移動後同じ部屋
12	午前	5–20 分	b	移動前と移動後同じビル、
13	午前	20–90 分		同じ階で違う部屋
21	昼休憩	0–5 分	c	移動前と移動後同じビルで
22	昼休憩	5–20 分		違う階
23	昼休憩	20–90 分	d	第 1 ビルと第 2 ビルの間の移動
31	午後	0–5 分	e	第 1 ビルと第 3 ビルの間の移動
32	午後	5–20 分	f	第 2 ビルと第 3 ビルの間の移動
33	午後	20–90 分		

4.3 NMMF によるパターン抽出

分析に用いる NMMF は [7] を参考にした。本節では入退室データを NMMF で因子分解した行列分解形、分析に用いる NMMF、因子行列の初期値の設定、因子数の決定方法と最後に抽出した因子行列について説明をする。

4.3.1 行列分解形

4.2 節で算出した入力行列 X、Y に NMMF を用いると図 2 のように因子分解できる。因子行列 A、B の積が入力行列 X

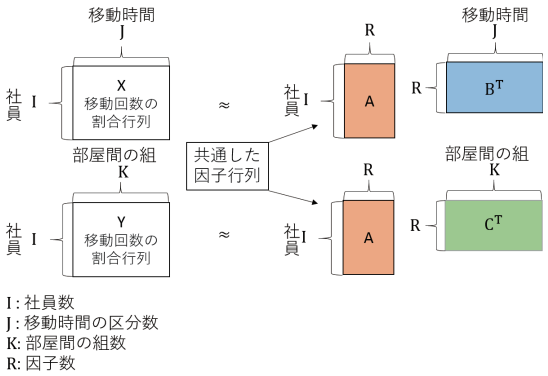

図 2: 行列分解形

の分解結果であり、因子行列 A と C の積が入力行列 Y の分解結果である。因子行列 A は各社員がどの因子にどの程度基づくのかを示す。因子行列 B は各因子の代表的な割合の多い移動時間、因子行列 C が各因子の代表的な割合の多い部屋間の組である。

4.3.2 定式化

図 2 の近似 ≈ の尺度には、ユークリッド距離を利用した。

$$d_{EU}(x_{ij}||\hat{x}_{ij}) = \frac{1}{2}(x_{ij} - \hat{x}_{ij})^2. \quad (1)$$

d は行列の要素同士の距離であり、行列同士の距離 $\mathcal{D}$ は式 (2) のように表せる。

$$\mathcal{D}_{EU}(X||\hat{X}) = \sum_{i,j=1}^{I,J} d_{EU}(x_{ij}||\hat{x}_{ij}). \quad (2)$$

ここで、入力行列 X と因子行列の積 $\hat{X}$ の距離を $\mathcal{D}(X||\hat{X})$（入力行列 Y も同様 $\mathcal{D}(Y||\hat{Y})$）とし、NMMF は式 (3) に示す最適化問題を解くことで因子行列を出力する。因子行列を更新していく上で入力行列と因子行列の積の誤差（距離）を損失関数の値とする。

$$\arg\min_{A,B,C}\{\mathcal{D}_{EU}(X||\hat{X}) + \mathcal{D}_{EU}(Y||\hat{Y})\} \quad s.t. A, B, C \geq 0. \quad (3)$$

この最適化問題を解くアルゴリズムは複数存在するが、ここでは実装上の簡易さから利用されることの多い式 (4)-(6) の乗法更新則に基づくアルゴリズムを利用した。

$$a_{ir} \leftarrow a_{ir} \frac{\sum_{j=1}^{J} x_{ij} b_{jr} + \sum_{k=1}^{K} y_{ik} c_{kr}}{\sum_{j=1}^{J} \hat{x}_{ij} b_{jr} + \sum_{k=1}^{K} \hat{y}_{ik} c_{kr}}. \quad (4)$$

$$b_{jr} \leftarrow b_{jr} \frac{\sum_{i=1}^{I} x_{ij} a_{ir}}{\sum_{i=1}^{I} \hat{x}_{ij} a_{ir}}. \quad (5)$$

$$c_{kr} \leftarrow c_{kr} \frac{\sum_{i=1}^{I} y_{ik} a_{ir}}{\sum_{i=1}^{I} \hat{y}_{ik} a_{ir}}. \quad (6)$$

4.3.3 因子行列の初期値の設定

入力行列 X,Y は要素が割合で表してあることからの各行の合計 1 になることが分かる。そこで、本論文では更新をスムーズに進めるために、因子行列の初期値を式 (7)-(9) に示すように設定した。

$$A \sim Uniform\left\{0, \frac{1}{\sqrt{J*R}} + \frac{1}{\sqrt{K*R}}\right\}. \quad (7)$$

$$B \sim Uniform\left\{0, \frac{2}{\sqrt{J*R}}\right\}. \quad (8)$$

$$C \sim Uniform\left\{0, \frac{2}{\sqrt{K*R}}\right\}. \quad (9)$$

この初期値は入力行列の各行の合計が 1 になることを考慮し、因子行列同士を掛け合わした行列の各行の合計が 1 に近くなるように設定している。これによって、入力行列に近い初期値を与えることが可能となる。

4.3.4 因子数の決定方法

因子行列は因子数を増やすほど損失関数の値は減少していくため、損失関数の値の減少幅が明らかに小さくなった場合、そこを最適な因子数とする [10]。入力行列 X,Y を因子分解した結果、因子数は損失関数の値の減少幅が明らかに下がった因子数 9 とする。因子分解により得られた 3 つの因子行列 A,B,C のヒートマップを図 3 に示す。色の濃いマス程値が大きく、色の薄いマス程値が小さくなるように表示している。

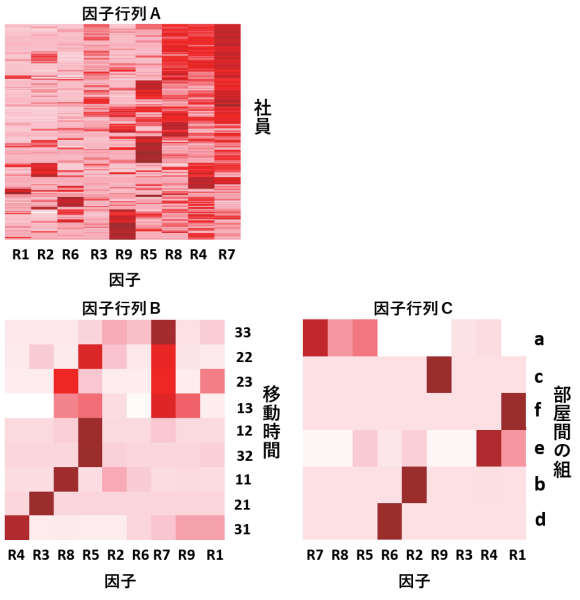

図 3: NMMF による因子分解の結果

表 2: クラスタ（グループ）内に含まれる社員数

クラスタ番号	1	2	3	4	5	6	7	8	9	10	11
社員数	26	14	8	13	11	5	46	29	21	51	108

4.4 因子行列によるクラスタリング

社員がどの因子にどの程度基づくのかを示す因子行列 A(図 3）の各行を社員の特徴量ベクトルとし、クラスタリングを用いた社員のグループ分けを行う。クラスタリングには k-means 法を用いて、クラスタ数は elbow 法を用いて決定した。結果クラスタ数 11 で社員のクラスタリングを行った。クラスタ中心のヒートマップを図 4 に、クラスタに含まれる社員数を表 2 に示す。

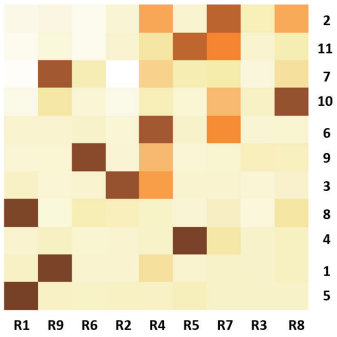

図 4: 因子行列 A をクラスタリングした結果のクラスタ中心のヒートマップ、行がクラスタ番号、列が因子

4.5 決定木学習を用いた分類木の作成

次に、クラスリング結果を用いて決定木学習を行う。決定木学習に用いる説明変数を社員属性（性別、部署、職種、採用種別、年代、社歴）とし、最大の階層を 3 とした。全クラスタの分類木の内、クラスタ人数の最も多いクラスタ 11 の結果を図 5 に示す。

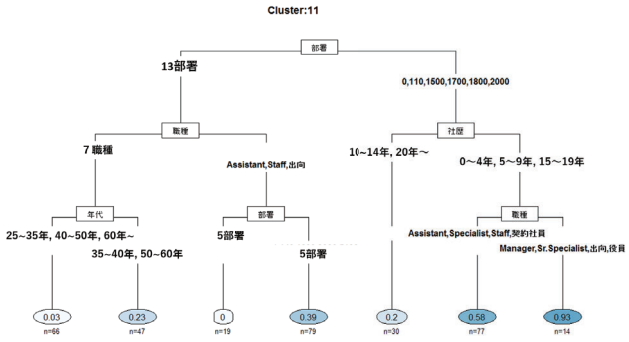

図 5: クラスタ 11 の分類木

4.6 考察

各クラスタがどの因子に当てはまるのかを図 4 及び因子行列 B と C（図 3）から特徴を見て、決定木学習により得られた分類ルールの内、カバー率の高いものから社員属性とパターンの関係を考察していく。特に社員属性の影響によりパターンが分かれていると考えられる結果のクラスタ 3、4、11 について考察する。クラスタ 3 とクラスタ 11 は社員属性の影響によりパターンが異なると考えられる結果である。またクラスタ 4 とクラスタ 11 は最も当てはまる因子は同じであるが、クラスタ 11 はさらに少し当てはまる因子を持っているため、クラスタ 4 との比較を行う。

- クラスタ 3 の 8 名の社員は主に因子 2 に当てはまり、因子 4 も少し当てはまっていた。因子 2 は、移動時間の割合が均一であることから移動時間がばらばらな因子であると読み取れる。また、移動には同じ階の違う部屋へ移動することが多かった。分類木はある一つの部署で職種が Manager、Sr. Specialist、Staff である 15 名が 27 %の確率でクラスタ 3 に分類されていた。以上から、この部署の役職の高い社員 (Manager、Sr. Specialist) と Staff は基本的には同じ階の会議室とオフィス（オフィス以外は会議室が多い）を往復することが多く、移動の途中でトイレ休憩や小休憩などを挟みながら仕事をしていると推測できる。
- クラスタ 4 の 13 名の社員は主に因子 5 に当てはまっていた。同様にクラスタ 11 も因子 5 に当てはまるが、クラスタ 4 とは違い因子 7 も少し当てはまっていた。因子 5 は午前は比較的に短い移動をしており、昼休憩と午後は 5–20 分の移動が多く、同じ部屋へ戻ることが多かった。分類木は、ある 3 つの部署で職種が Assistant、Manager、Specialist の年代が 25–30 歳、35–40 歳,、50–55 歳の 10 名が 40 %の確率でクラスタ 4 に分類されていた。以上から、この三つの部署の 25–30 歳、35–40 歳、50–55 歳の役職の高い社員（Manager、Specialist）はクラスタ 3 の役職の高い社員とは異なり、基本は自分のオフィスで仕事をしており、午前はトイレ休憩など移動に時間をかけていない。しかし、午後になるに連れ、小休憩をとっている（少し長めの移動（5-20 分）をしていることから）と推測できる。
- クラスタ 11 の 108 名の社員もクラスタ 4 と同様に因子 5 に当てはまり、因子 7 も少し当てはまっていた。因子 5 はクラスタ 4 にて示した通りの特徴を持つ。因子 7 は午前と午後に 20-90 分の移動をする割合が多く、昼休憩では比較的に移動に時間をかけており、同じ部屋に戻ってくる移動が多かった。分類木は、ある 6 つの部署で社歴が–9 年もしくは 15–19 年の職種が Manager、Sr. Specialist、出向、役員の 14 名が 93 %の確率でクラスタ 11 に分類されていた。以上から、これらの部署の役職の高い社員（Manager、Sr. Specialist、役員）もクラスタ 4 の社員と同様な仕事のスタイルであると考えられるが、因子 7 の特徴も持っていることから、たまに社外で会議があったり、昼休憩にランチをとり行くことがあると推測できる。

以上より、役職が同じ場合でも部署で行動パターンが異なることや、クラスタ中心を各クラスタの特徴ととらえることでクラスタ 4 とクラスタ 11 の比較のように、一番当てはまる因子は同じ場合でも、少し異なる特徴を持っていることを発見できることがわかった。その他にも、ビル間の移動に時間をかけない社員など様々な社員属性とパターンの関係がみられた。

5. まとめと今後の方針

本研究では NMMF により抽出されたパターンと属性情報の関係を分析する手法を提案した。提案手法を用いてオフィスの入退室データを用いて社員の移動時間と部屋間の組に関する行動パターンと社員属性の分析を行ったところ、入退室データの分析により、役職が同じ場合でも部署で行動パターンが異なるなど、提案手法を用いて属性情報とＮＭＭＦにより得られたパターンの関係を分析することの有効性を確認した。しかし、クラスタ数の決定方法に elbow 法を用いたが、分析結果の中で同じ特徴を持ったクラスタがあった為、最適なクラスタ数とは言い切れないと考える。

今後の方針としては、クラスタ数の決定方法の見直しや決定木学習の最大の階層の指定方法などの課題が挙げられる。

謝辞

本研究を進めるにあたり、入退室データを提供して頂いた協力企業に感謝の意を表する。本研究は JSPS 科研費 JP18K18160 の助成を受けたものです。

参考文献

[1] 嶋本寛, 北脇徹, 宇野伸宏, 中村俊之, “IC カード利用履歴データを用いた公共交通需要変動分析”, 土木学会論文集 D3(土木計画学), Vol.70, NO.5(土木計画学研究・論文集第 31 巻) , pp.605-610, 2014.

[2] 鈴木敬, 相薗敏子, “交通 IC カード利用履歴を用いた生活行動属性指標の提案”, 信学技報, IEICE, Technical Report, LOIS2011-84, 2012.

[3] 佐藤雅之, 及川和彦, 永嶋規充,“入退室管理システムにおける通行履歴の応用”, 情報処理学会第 77 回全国大会, 2G-05.

[4] 岡田将吾, 神谷祐樹, 佐藤祐作, 藤田義弘, 山田敬嗣, 新田克己,“センサ環境を利用したオフィスワーカーの行動パターン分析”, 第 27 回人工知能全国大会, 1C4-1in, 2013.

[5] Seidai Kojima, Hayato Ishigure, Miwa Sakata, Atsuko Mutoh, Koichi Moriyama and Nobuhiro Inuzuka,“An Analysis Method of Traveling-Time Patterns Between Rooms from Entry and Exit Data of Office Workers”, 2018 IEEE 7th Global Conference on Consumer Electronics(GCCE2018), pp.267-270.

[6] 小島世大, 石樽隼人, 坂田美和, 武藤敦子, 森山甲一, 犬塚信博,“オフィスワーカーの入退室データを用いた移動時間パターンの分析”, 第 32 回人工知能全国大会, 302-OS-1b-01, 2018.

[7] 幸島匡宏, 松林達史, 澤田宏 “複合データ分析技術と NTF［1］――複合データ分析技術とその発展――”, 電子情報通信学会誌 Vol. 99, No. 6, 2016.

[8] 幸島匡宏, 松林達志, 澤田宏 “属性情報を考慮した消費者行動パターン抽出のための非負値多重行列因子分解法”, 人工知能学会論文誌 30 巻 6 号 SPI-G, 2015.

[9] 古川康一, 升田俊樹, 西山武繁 “チェロ演奏動画の目視によるデータ獲得と演奏スタイルの分類”, 第 30 回人工知能学会, 1M4-OS-14a-3, 2016.

[10] 安川武彦 “非負値行列因子分解を用いたテキストデータ解析”, 計算機統計学, 第 28 巻・第 1 号: 2015, pp.41-55.