CL-Machine LearningにおけるNMFパッケージについて

2010年5月7日

知識工学部　阿部裕介

発表概要

1. CL-Machine Learning(CLML)とは?
2. NMFについて
3. NMFの実装について
4. NMFの適用例(スポーツ・コーパスデータの分析)

1. CLMLとは?

知識工学部で開発している

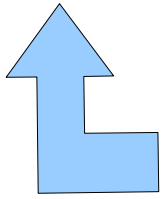
Common Lispで書かれた

統計・機械学習(Machine Learning)

パッケージ

CLMLの目指すもの

- High Performance
- Large Scale



昨今ではマルチコア&CPUが主流になりつつあり、
その性能を活用するために、CLMLでは並列処理を
強く意識している

CLMLの特長

- プラットフォーム非依存
- fork-futureによる並列計算
- Intel MKLによる高速な線形代数演算

◇さらに並列処理を直感的に行える
データフロービジュアル
プログラミング環境も追加予定！

2. NMFについて

- NMF = Non-negative Matrix Factorization
  = 非負行列因子分解

- 比較的最近提案された(Lee and Seung, 1999)
  教師なし学習法

- PCA(Principal Components Analysis)の
  代替手法として注目されている

NMFの前提

- NMFが行うのは、非負行列として表現されたデータセットXを2つの非負行列WとHの積に近似分解すること

$$\Rightarrow X \simeq WH$$

- NMFを用いてデータの分析を行う上での前提は、分析にかけたいデータセットを、何らかの特徴要素の非負線形和として捉える見方が妥当であること

NMFができると何が嬉しいのか

- 次元縮約

  例えば、テキストデータの行列表現は

  一般に高次元、スパースで扱いづらい

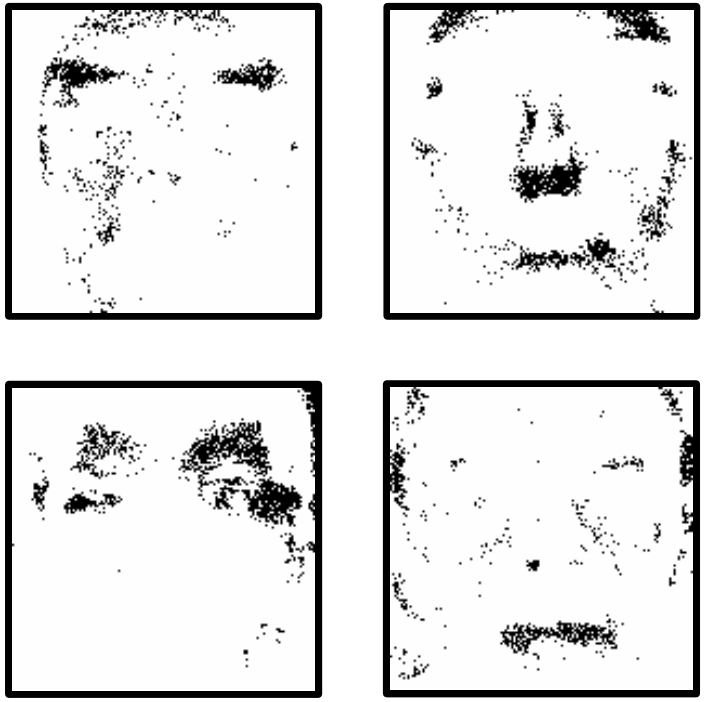
  NMFで次元を下げてから、ベクトル空間モデルを作成

- 特徴抽出

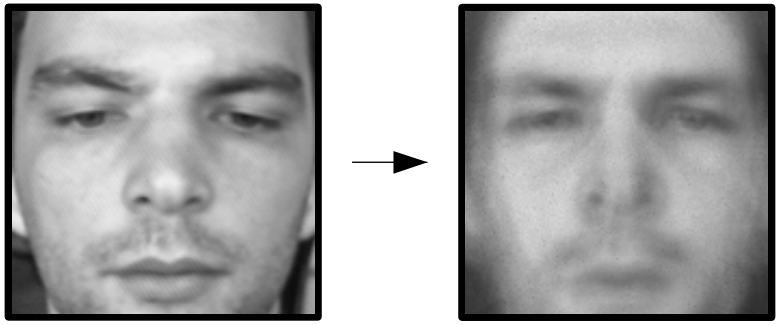
  データセットに内在する「特徴」を発見する

  有名な例としては顔の構成パーツ（目・鼻）の抽出等

NMFによる顔パーツ抽出例



NMFによって抽出された顔の特徴要素



オリジナル画像

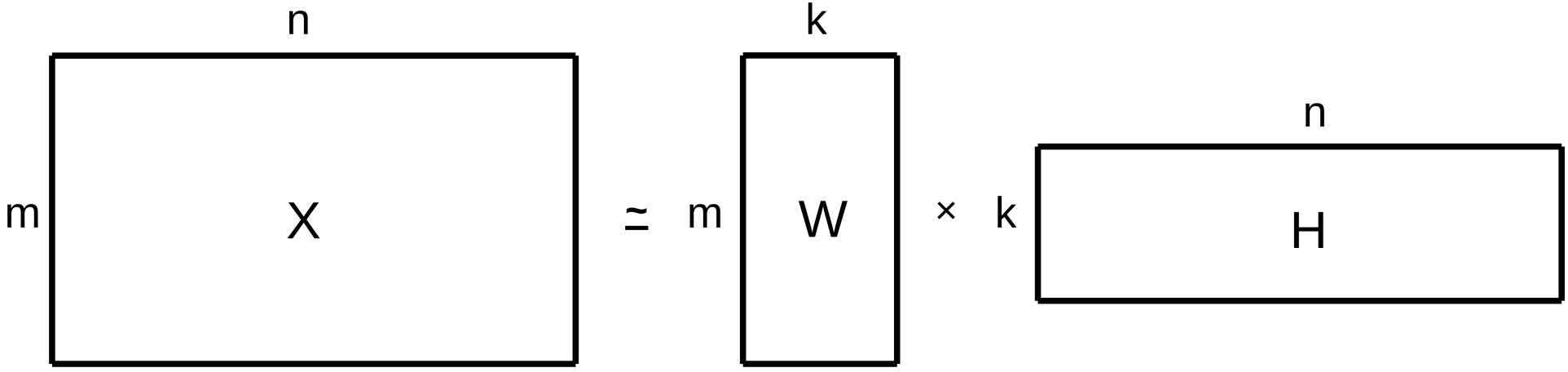
次元縮約した後
再構成した画像

NMF(1)

- X : (m,n)-行列
  W : (m,k)-行列
  H : (k,n)-行列

$$X \simeq WH$$

k次元に縮約
k個の特徴を抽出



NMF(2)

- W,Hの初期値はランダムな非負値で埋めておくのが一般的(前処理のクラスタリング結果を初期値に用いる手法もあり)

- 基本的なアイディア

  「XとWHのノルムの差を最小化する最適化問題」

NMFの乗法的更新アルゴリズム(1)

- 最小化

$$||X - WH||^2$$

- 制約条件

$$W_{ij} \geq 0 \ (1 \leq i \leq m, 1 \leq j \leq k),$$
$$H_{ij} \geq 0 \ (1 \leq i \leq k, 1 \leq j \leq n)$$

NMFの乗法的更新アルゴリズム(2)

- 乗法的更新式(Lee and Seung, 1999)

$$W_{ij} \leftarrow \frac{(XH^T)_{ij}}{(WHH^T)_{ij}} W_{ij},$$

$$H_{ij} \leftarrow \frac{(W^T X)_{ij}}{(W^T WH)_{ij}} H_{ij}$$

※ 但し、局所最適解に収束

機械学習と最適化問題

- 機械学習でわりとよく見られるパターンは、
  ある種の最適化問題に帰着させて、
  それを固有の方法で効率的に解くというもの

- 他の例では、
  Support Vector MachineにおけるSMOアルゴリズム

3. NMFの実装について

- 開発期間

  2009年9月から10月にかけて

今まで手がけた機械学習パッケージ

NMF　9~10月

階層型クラスタリング　9~10月

→ 教師なし（分析的）機械学習

決定木　11月

Random Forest　12月

SVM　1~4月

→ 教師あり（予測的）機械学習

※2~3月はNII(国立情報学研究所)の機械学習を用いた案件に従事
CLMLはそこでも開発基盤として利用されている

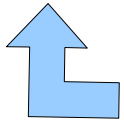
Common Lispの良い点

- 速やかにプロトタイプを作成し、
  それを効率的にチューニングしていける
- 対話的開発環境のため、テスト・デバッグが
  非常にやりやすい
- 関数的にも手続き的にも書ける強み
- 時の試練を経ていて、安定した仕様を持つ

実装で苦労した点

- パフォーマンス・チューニング

  初期バージョンは破天荒に遅かった…



実際には、黒田さん、黄さんに
チューニングをしていただいた

今後は自分でもチューニング
できるようにしたい

パフォーマンスチューニング結果

```
NMF(4): (setf x (sample-matrix 100 1000))
#2A((64.0 62.0 15.0 80.0 47.0 33.0 76.0 20.0 75.0 12.0 ...)
    (42.0 28.0 66.0 83.0 52.0 51.0 13.0 43.0 48.0 18.0 ...)
    (57.0 93.0 87.0 95.0 31.0 86.0 58.0 10.0 57.0 53.0 ...)
    (29.0 38.0 56.0 49.0 59.0 74.0 14.0 71.0 58.0 71.0 ...)
    (91.0 47.0 33.0 7.0 2.0 1.0 83.0 1.0 33.0 64.0 ...)
    (82.0 99.0 21.0 89.0 24.0 53.0 15.0 27.0 76.0 80.0 ...)
    (88.0 54.0 89.0 64.0 69.0 20.0 16.0 65.0 80.0 41.0 ...)
    (87.0 50.0 15.0 45.0 28.0 42.0 43.0 22.0 93.0 69.0 ...)
    (72.0 7.0 97.0 84.0 33.0 82.0 50.0 96.0 94.0 58.0 ...)
    (93.0 86.0 0.0 72.0 91.0 85.0 4.0 6.0 72.0 66.0 ...)
    ...)
NMF(5): (time (old-nmf::nmf-gamma x 5))
; cpu time (non-gc) 25,890 msec user, 250 msec system
; cpu time (gc)      12,360 msec user, 20 msec system
; cpu time (total)   38,250 msec user, 270 msec system
; real time  38,803 msec
; space allocation:
;  1,190 cons cells, 26,159,703,792 other bytes, 0 static bytes

NMF(7): (time (nmf x 5))
; cpu time (non-gc) 1,370 msec user, 0 msec system
; cpu time (gc)      0 msec user, 0 msec system
; cpu time (total)   1,370 msec user, 0 msec system
; real time  1,371 msec
; space allocation:
;  1,024 cons cells, 5,263,792 other bytes, 0 static bytes

NMF(10): (time (mkl-vector::mkl-nmf x 5))
; cpu time (non-gc) 940 msec user, 0 msec system
; cpu time (gc)      10 msec user, 0 msec system
; cpu time (total)   950 msec user, 0 msec system
; real time  252 msec
; space allocation:
;  13,208 cons cells, 186,639,392 other bytes, 2,400 static bytes
```

28.3倍

5.4倍

トータルで
153倍！

4. NMFの適用例(スポーツ・コーパスデータの分析)

- 分析対象

  東京読売のスポーツ記事100本

  形態素解析をして1202語を抽出

BOW表現

- 先のスポーツ・コーパスデータを
  NMFで分析するために、まずはコーパスを
  非負行列として表現することが必要

- BOW = Bag of Words
  ナイーブには、行に記事を、列に単語をとり、
  成分として単語の出現頻度を持つ行列

Weighting Schemeの重要性

- The success or failure of the vector space method depends on the term weighting schemes

  (1992,Erica Chisholm and Tamara G. Kolda,

  New Term Weighting Formulas for the Vector Space Method in Information Retrieval)

- データの元の素性がもっともよく反映・保持

  されるように、n次元ベクトルとして

  データを表現することが機械学習では肝要

TF-IDF

- TF-IDF : Weighting Schemeの代表例

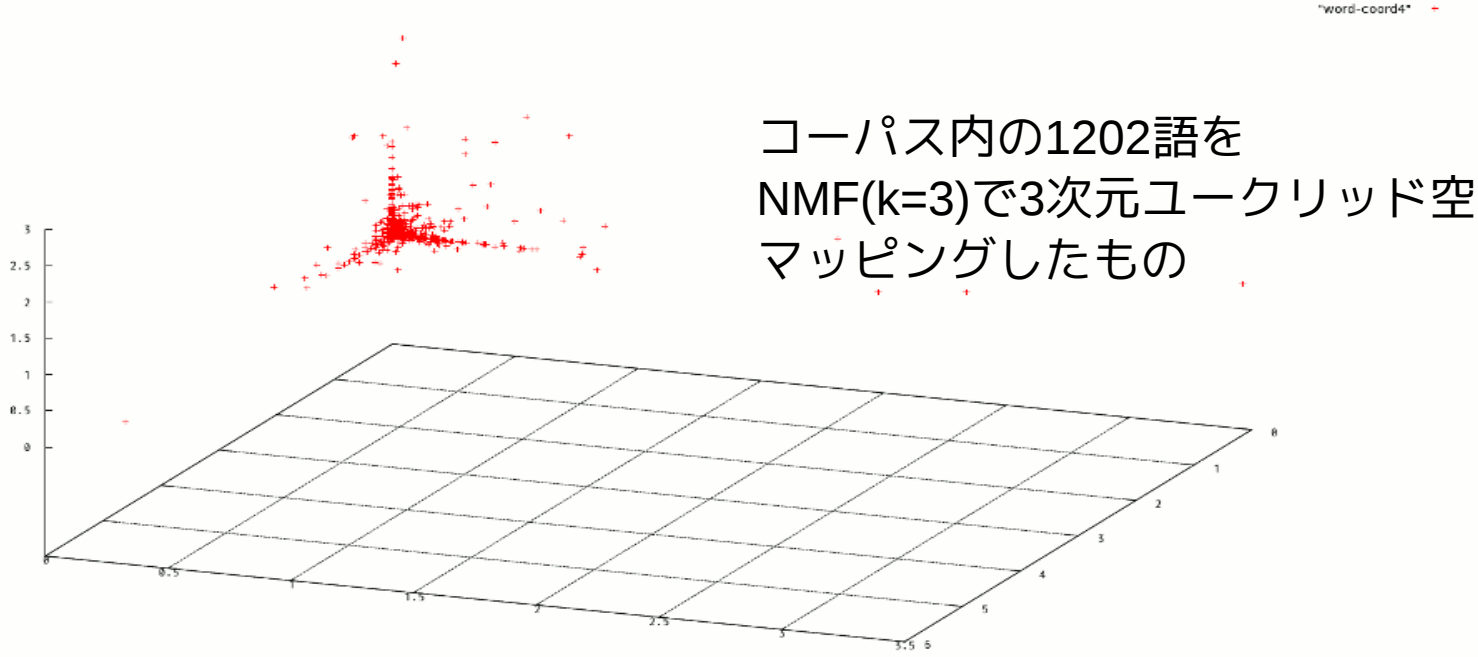
  Term Frequency-Inverse Document Frequency

- Weighting Scheme如何で抽出されてくる特徴や形成されるクラスタは大きく異なる

  （例えば、長い文書が強い影響力を持つ、など）

NMFによるコーパス特徴抽出

- 行列表現されたコーパスデータを
  k=3でNMFにかけると、3つの特徴成分の
  強度に応じて、文書・語が配置される



コーパス内の1202語を
NMF(k=3)で3次元ユークリッド空間に
マッピングしたもの

NMFによるコーパスの分析(1)

・語と文書が同時に、同種の特徴にしたがって、スコア付けされる

```
NMF(10): (nmf-corpus-analysis sports-corpus 3 :results 5)

Feature 0
マラソン      0.08301250336134758
大阪      0.05615447060365708
世界      0.053483632442399384
練習      0.03968124799189396
日本      0.03584383586221997

Feature 1
決勝      0.1238922779961146
成年      0.1161866648781798
少年      0.10784310336215523
アイスホッケー      0.09516233144069992
男子      0.09352080674129722

Feature 2
キャンプ      0.09385654379913695
監督      0.09154990021501706
宮崎      0.08516018488895362
投手      0.0665598294202915
野村      0.06102378318229685

Feature 0
00267800      2.8134678237937365
00261590      2.595820348134457
00267780      2.5429335848291412
00267810      2.414408689897669
00267690      1.7478149127725302

Feature 1
00264720      2.060530605198221
00265130      2.018462327125502
00264810      1.9041283857681675
00265920      1.7244743077622926
00265250      1.682330628208047

Feature 2
00260660      1.4374873686685459
00264500      1.358756940758196
00261710      1.3099241114639195
00260650      1.2698169338729624
00261770      1.1067646081307896
NIL
```

Feature2は「野球」という特徴を抽出している、と推測できる

NMFによるコーパスの分析(2)

- Feature2成分がもっとも強かった

  記事00260660を実際に見てみると…

２月１日からの春季キャンプに向け、巨人の原監督らが３１日、空路で宮崎入りした。
既に合同自主トレでキャンプ一軍メンバー３３選手が宮崎に入っており、
この日、首脳陣と外国人選手ら３２人が宮崎空港に到着。歓迎セレモニーで、原監督は「奪回」を改めてテーマとして示し、
「体力、技術、英気を養うために宮崎で練習していく。我々の戦いざまを見てほしい。日本一目指して頑張っていきます」とファンに誓った。
原監督らはその後、宮崎神宮などでキャンプの無事とシーズンでの勝利を祈願した。
キャンプは一軍が２５日まで、二軍は２６日までの日程で、宮崎県総合運動公園で行われる。

写真＝宮崎空港に到着し、歓迎の花束を受け取る巨人・原監督（左）と清武代表（午前１１時３０分）

NMFによるコーパスの分析(3)

- Feature2成分が次に強かった記事00264500

◇行くぞ！イーグルス
◆「一段一段頂点目指す」
沖縄・久米島で行われるキャンプを前に、野村克也監督率いる楽天イーグルスの
選手たちが３０日、仙台市青葉区の大崎八幡宮で必勝祈願した。

集まったのは、野村監督をはじめ、選手、コーチや球団スタッフら総勢約１００人。
野村監督が、キャンプ前の恒例行事として、チーム全体で必勝祈願することを提案した。
おそろいのスーツに身を包んだ選手たちは、「畳一畳分」ほどの特製の絵馬に
それぞれサインを書き込んだ後、神社の本殿でお払いを受け、必勝祈願した。
絵馬には野村監督が最後に筆を入れた＝写真＝。

NMFの問題点(1)

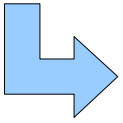
- kの決定

  データセットに内在する自然な特徴の個数kが

  そもそも一意にあるか、あるとして

  どのようにそれを決定することができるか、

  という問題



注目すべき新手法として、
ノンパラメトリックベイズによる
「kの事前設定を必要としない」分析がある
(CLMLでは今年4月に実装済)

Chinese Restaurant Process

- 「中国の人は混んでいるテーブルを好む」

  テーブルの数 = クラスタ数 = k

  データをひとつずつ追加していき、

  新規データは小さい確率で新しいクラスタを形成する

  （そうでない場合、既存の混んでいるテーブルに着座）

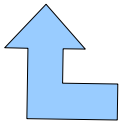
NMFの問題点(2)

- 局所最適解に収束

  大域最適解への収束は未保証

  W,Hは一般にはランダムな初期値から

  出発するため、分解結果が毎回異なりうる



すなわち、異なる「特徴」が抽出されてくる

「重み付け」による特徴誘導

- NMFが局所最適解に収束し、毎回異なる特徴が抽出されてきてしまう問題への対処のひとつとして、「重み付け」による特徴誘導が挙げられる

- 具体的には、こちらが「テーマ・トピック」として設定したい単語または記事の「重み」を上げて、コーパスデータを非負行列化し、NMFにかける

特徴誘導を利用した検索

- 「西武」という単語の重みを上げて、

  k = 1でNMFにかけると、以下のような結果が得られる

```
NMF(12): (nmf-corpus-search sports-corpus "西武" :type 'term :results 5)

Feature 0
西武      0.5021197972696246
所沢      0.030414168851152966
埼玉      0.03008766396664986
期待      0.028342049167192684
松坂      0.024213725076168025

Feature 0
00261790      9.384323367473277
00266250      4.432653799514354
00261710      1.348321581690453
00261730      0.09358571006030018
00265240      0.06357081136906965
```

「西武」という語に関連する単語や記事が、その関連度のスコア付きで得られる

「西武」という語を含んでいないが、関連性の高い記事をコーパス内から抽出

記事00261730

ソフトバンクの新外国人４選手が３０日、ヤフードームで入団記者会見を行い、抱負を語った。

ヤクルトでの２年間で１７勝を挙げたガトームソン投手（３０）は、「ベストを尽くして優勝に貢献したい」。２年総額２億５０００万円で契約し、背番号は「４３」に決まった。

他の３選手は、それぞれ単年の年俸５０００万円で契約。ニコースキー投手（３３）（元ナショナルズ）は背番号「３５」、アダム外野手（３５）（元レンジャーズ）は「４」、ブキャナン内野手（３３）（元パドレス）は「００」に決まった。（金額は推定）

ご清聴ありがとうございました