機械学習によるメンタルヘルス・マネジメント支援

2010年7月28日

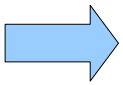
知識工学部　阿部裕介

概要

1. 問題定義
2. 決定木によるアプローチ
3. トピック抽出 × HMM

1. 問題定義

- 社員の配属先・勤務状況・健康診断結果などから
  精神健康状態を推定・分析したい
- さらには社員の日報・Blog・Twitterを解析し、
  精神健康状態の推定に役立てたい



決定木やトピック抽出、隠れマルコフモデル（HMM）といった機械学習システムによるメンタルヘルス・マネジメント支援

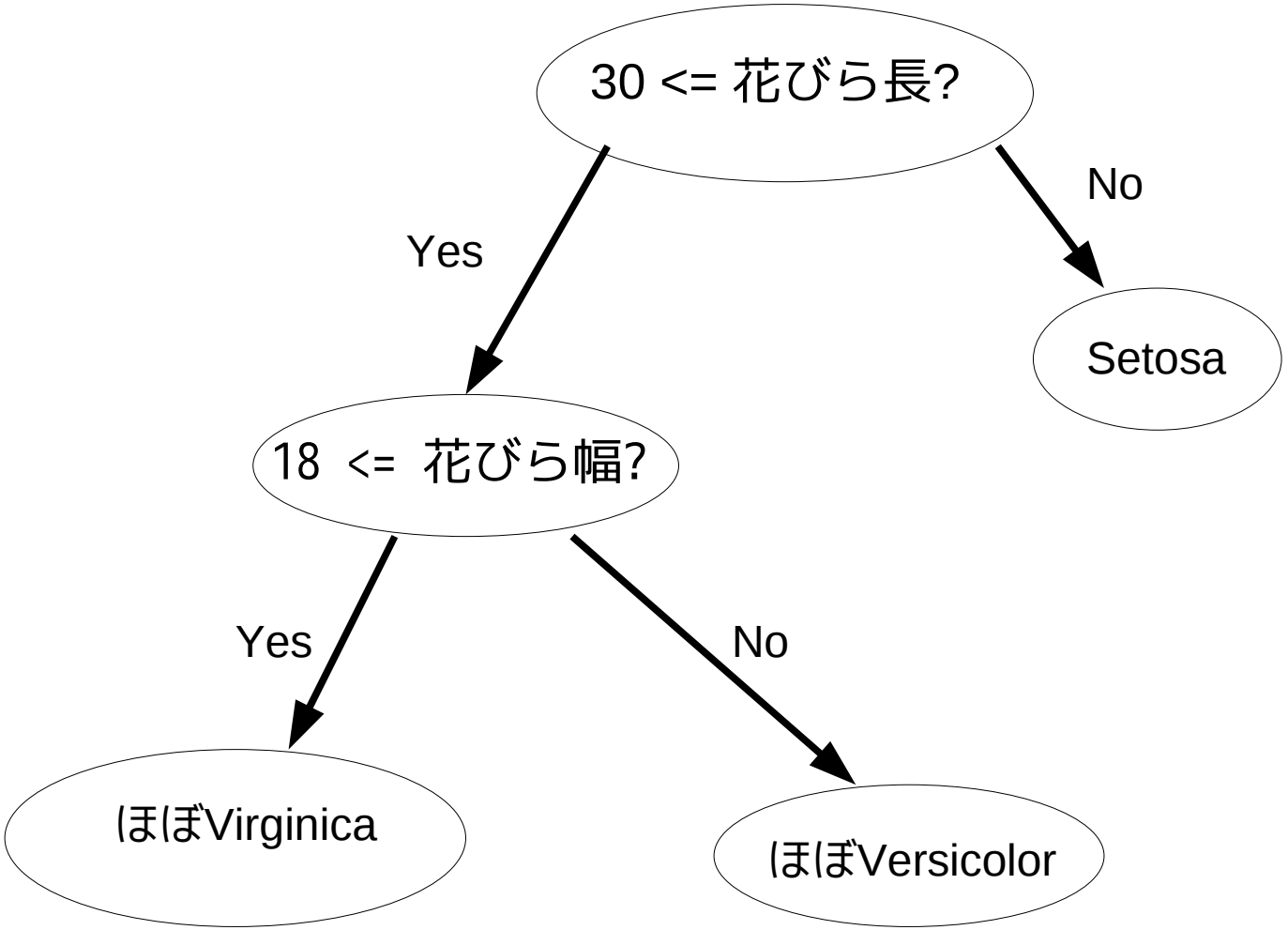
2. 決定木によるアプローチ

- 機械学習としては簡単かつ古典的な手法だが、実際的な方法として、決定木によるアプローチをまずは紹介する
- 精神健康状態の推定だけならばSVMでも可能.しかし、決定木では推論過程が可視的であり、危険因子の特定や疾病に至る共通パターン発見といった、分析的用途にも使うことができる

決定木の例（Fisher's Iris）

```
[30 <= 花びら長?]((Virginica . 50) (Versicolor . 50) (Setosa . 50))
   Yes->[18 <= 花びら幅?]((Versicolor . 50) (Virginica . 50))
      Yes->((Virginica . 45) (Versicolor . 1))
      No->((Virginica . 5) (Versicolor . 49))
   No->((Setosa . 50))
```

- 「花びら長」が30より小⇒Setosa(50/50)
- 「花びら長」が30以上、かつ
  「花びら幅」が18以上⇒ほぼVirginica(45/46)
- 「花びら長」が30以上、かつ
  「花びら幅」が18より小⇒ほぼVersicolor(49/54)


30 <= 花びら長?
No
Yes
Setosa
18 <= 花びら幅?
Yes
No
ほぼVirginica
ほぼVersicolor

決定木の原理

- 先の例でみたような、Yes/No分岐ルール群を学習用データセットから自動的に生成

- 具体的には判別結果を「もっともすっきり」わける述語を都度選択して、木をつくっている

ジニ不純度やエントロピーを用いている

決定木による精神健康状態の推定

- 性別・年齢・所属・配置転換の有無・睡眠状態などから特徴素ベクトルを作成

⇒ どのような特徴素を選ぶかが決定木分析の鍵
決定木のメリットとしては、数値データと
カテゴリカルデータが混在しているデータでも
そのまま分析にかけることができる

- 得られた特徴素ベクトルと精神健康状態を示すクラスラベルを用いて決定木を作成（学習）

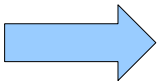
⇒ 得られた決定木を推定に用いる

決定木による精神健康状態の分析

- 学習によって得られた決定木の推論過程を観察することで、社内において特定の精神疾患に陥る経路の発見などが期待できる

- 決定木の発展形としてRandom Forest がある.

  その枠組みではどの特徴素がもっとも推定に寄与しているかを定量的に評価できる



Variable Importance

3. トピック抽出 × HMM

- ここでは決定木のような古典的手法ではなく、NMFやHDP-LDAといった新しい機械学習手法を利用したシステム案について紹介する

- 社内日報・Blog・Twitterといったテキストデータからトピックを抽出し、それを観測可能な変数として個々の精神状態毎に確率モデル(HMM)をつくり、社員の精神健康状態を推定する

cf. 若林・三浦「HMMを用いた文書における事象系列の推定」

トピック抽出

- 非負行列因子分解（NMF）やHDP-LDAといった機械学習手法を用いることで、テキストデータからトピックを抽出することができる

（毎日jpのニュースアーカイブ5/7~7/6までの政治記事をNMFでトピック抽出した事例）

Feature 3
消費税　0.017450445481965633
首相　0.01177200312731532
マニフェスト　0.011181456967734362
議論　0.01091955694593518 5
社会保障　0.010533312900445686
財政　0.009098150329703108
財源　0.00894970685916 5055
増税　0.0087706671074 6533
消費税増税　0.008741467092357912
国民　0.00872065786647 0782

Feature 7
沖縄　0.022772689725362158
負担軽減　0.013925261452755685
日米共同声明　0.010433503021361194
理解　0.009142491165829688
沖縄県　0.009120778247201651
米軍普天間飛行場　0.008969987846265706
工法　0.00823401072543 6105
沖縄県宜野湾市　0.007567833390501 9645
仲井真知事　0.007186171295733188
会談　0.007158637617708451

HDP-LDAによる介護Blogからのトピック抽出例

(("Topic 1" #("母" "小規模ホーム" "右膝" "状態" "自分" "家" "介護" "ヘルパーさん" "人" "テーピング") . 0.0121729737259825 05)
("Topic 8" #("痰" "抗生剤" "噎せ" "発熱" "看護師" "主治医" "咳" "発生" "肺炎" "対処") . 0.0093561007650118 72)
("Topic 2" #("施設" "利用者" "ケアマネさん" "夏みかん" "変更" "利用" "介護施設" "家族" "スタッフ" "泊まり") . 0.0090011398039457 49)
("Topic 7" #("主治医" "病院" "痛み" "整形外科" "軽減" "負担" "水" "総合病院" "湿布" "レントゲン") . 0.0088500828453500 66)
("Topic 11" #("歯医者" "入れ歯" "歯" "準備" "治療" "根本的" "先生" "良医" "医師" "認知症") . 0.0085969778632147 03)
("Topic 6" #("軸足" "体" "右足" "位置" "上着" "方法" "理学療法士" "肩" "移乗時" "移乗") . 0.0085930602118304 7)
("Topic 4" #("業者" "側溝" "お隣さん" "工事" "修復" "報告" "掃除" "謝り" "介助者" "脱輪") . 0.0085927911848085 13)
("Topic 3" #("尿" "シャワーキャリー" "採取" "処理" "便" "便器" "居間" "タオル" "坊主" "だめ") . 0.0085577488773684 32)
("Topic 5" #("野菜" "ポトフ" "業者さん" "料理" "福祉用具" "カレー粥" "朝食" "定番" "車椅子" "冷凍") . 0.0085318441507561 65)
("Topic 49" #("仕事" "面接" "通い" "両立" "就職" "自分" "制限" "定員オーバー" "フロア" "模索") . 0.0085313553346348 42))

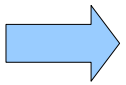
隠れマルコフモデル(HMM)

- 内部状態は単純マルコフ過程に従って遷移
  - 時刻tの状態のみに依存して、時刻t+1の状態が決定
- 各内部状態は観測可能なシンボルをひとつ、確率的に出力する
- 音声認識等において広く使われる確率モデル

トピック抽出とHMMを組み合わせて
精神健康状態の推定ができないか

若林,三浦「HMMを用いた文書における事象系列の推定」

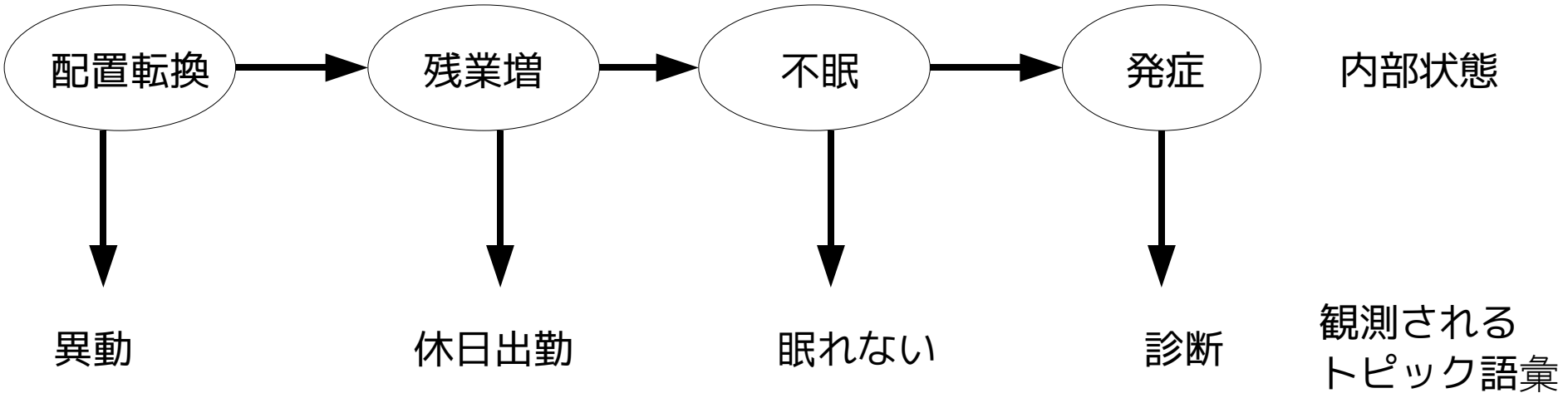
- 事象系列…事象(event)の時系列(sequence)
- 文書を事象系列とみなして分類したい
- この論文では、新聞の記事の中から単独犯事件・組織犯事件・汚職事件の３種類を異なる事象系列としてとりあげ、HMMを用いて推定・分類を行っている（平均正解率63.1%）



上記論文のアナロジーとして、
精神健康状態の時系列変化をHMMでモデリングし、
そのタイプの推定・分類問題を考える

イメージ

- うつ病の隠れマルコフモデル



※　このようなモデルを各精神健康状態毎につくり、下段の観測される
トピック語彙がもっともよく生成されるモデルを精神健康状態の推定結果とする

問題点

- 精神健康状態の遷移が単純マルコフ過程に従うという仮定はやや乱暴で、精神疾患は離れた時間間隔、複合的な要因で発生しうる

- トピック語彙は、身体的・精神的状態を示す語彙に制限する必要がある