7ZB-02

PCA と GMM を用いた血液検査データからの保健師指導効果の可視化

河田　昂 †　　大枝　真一 ‡
木更津工業高等専門学校　制御・情報システム工学専攻 †　　木更津工業高等専門学校　情報工学科 ‡

1. はじめに

近年，オープンデータが注目集めている．オープンデータとは，国や政府，地方自治体が保有する地理空間情報や防災，統計情報といった公共性の高いデータで，主に，信頼性向上，官民協働推進，経済活性化を目的として行われている活動のことである．現状では，地域によってその活動に大きな差があり，まだまだ浸透しているとは言い難い状況である．本研究では，木更津市からデータを提供していただき，オープンデータ活用の一例を示すものである．具体的には，通常十数項目存在する血液検査の項目を主成分分析を利用することによって二次元グラフ上にプロットし，年度毎の変化をアニメーションにすることによって，個人個人の変化を可視化し，厚生労働省によって定められているメタボリックシンドロームの階層化基準に合わせて個人を色分けすることによって，主成分分析後のグラフ上での健康な者と不健康な者のおおよそ位置を明確にすることにより，保健師の指導効果の可視化を行った．

2. 混合正規分布モデル

混合分布とは, 複数の確率分布を加重平均によって組み合わせたもので, 単純な分布を混合してより複雑なモデルを記述するためのモデルである. 混合分布は単純な構造でありながら柔軟な近似能力を持つことや, 比較的少数の組み合わせであっても非常に多様な分布を表現できることがメリットして挙げられる. データが与えられた時に, そのデータが各要素モデルから生成された可能性を事後確率として計算することによって, 教師なしデータのクラスタリングを行うことができる. モデルのパラメータを推定する手法としては EM アルゴリズムが用いられる.EM アルゴリズムは尤度関数の近似値を効率よく求めるための手法として使われる. 実際に用いられる要素モデルの代表的なものは正規分布であり, これを用いた混合分布モデルを混合正規分布モデル (Gaussian Mixture Model, GMM) と呼ぶ.

3. 主成分分析

変数が少ない場合は, 簡単なグラフや基本統計量などを利用することでデータの構造を明らかにすることができるが, 変数が多くなるとデータの構造が複雑になり, 解析が難しくなる. 一方で, 変数が多くなるとその間には相関がある可能性が高くなる. 主成分分析 (PCA: Principal Component Analysis) は, 多くの変数により記述された量的データの変数間の相関を排除し, できるだけ少ない情報の損失で, 少数個の無相関である合成変数に縮約して, 分析を行う手法である.

4. 計算機実験

4.1 データセット

本実験で利用したデータは木更津市役所に提供していただいたデータで木更津市民の血液検査結果のデータである. このデータは年齢, たばこ喫煙の有無, 階層化結果, 血液検査結果などの項目より成り立っている. 階層化結果は, 厚生労働省により定められているメタボリックシンドロームの基準を表すラベルである. 情報提供レベル, 動機づけ支援レベル, 積極的支援レベルが存在し, 後者のほうがリスクが大きい. 情報提供レベルではないと判断されたものであっても薬を処方されていれば, 情報提供レベルとしてラベリングされているため, 必ずしも情報提供レベルであればリスクが少ないと判断することはできない. 本実験では, 可視化を行うにあたって情報提供レベルの者と動機づけ支援レベルの者, 積極的支援レベルの者のグラフ上における比較を行うことによって, 保健師の指導効果の把握を行おうと考えているため, 本実験での情報提供レベルの者は薬を処方されていない元々情報提供レベルだった者のデータのみを利用している. 今回の研究に使用させて頂いたデータ数は 2012 年度から 2014 年度までそれぞれ 2510 名のデータを用いている.

Visualization an effectiveness of health care service from blood test data with PCA and GMM
†Akira Kawata・Advanced DJ Engineering Course, National Institute of Technology, Kisarazu College
‡Shinichi OEDA・National Institute of Technology, Kisarazu College

4.2 実験手順

1. PCA を利用して, 多次元な血液検査のデータを 2 次元に変換する.
2. GMM を利用して, 多すぎる被験者の数をグループ化し, 代表的な点をいくつか決定する.
3. 2012 年度から 2014 年度までの結果を 2 次元グラフ上にプロットする.
4. 2012 年度から 2013 年度,2013 年度から 2014 年度への各被験者グループの動きを可視化するために, 変化を細分化し, アニメーション形式に変換する.

5. 実験結果

以下の通り可視化をすることに成功した. これらの結果の点一つ一つが被験者を表しており,「情報提供:+」「動機付け支援:○」「積極的支援:*」として表現している. これらのグラフ結果を年度毎にアニメーションにすることによって可視化した. 例えば, 2012 年度から 2013 年度までのアニメーションは, 被験者それぞれに対する位置の差分を取り直線的に 30 分割して移動させることによりアニメーションとした. 情報提供者に対する動機づけ支援者と積極的支援者の位置を相対的に比較することにより保健師の指導効果の把握を行うことができる.

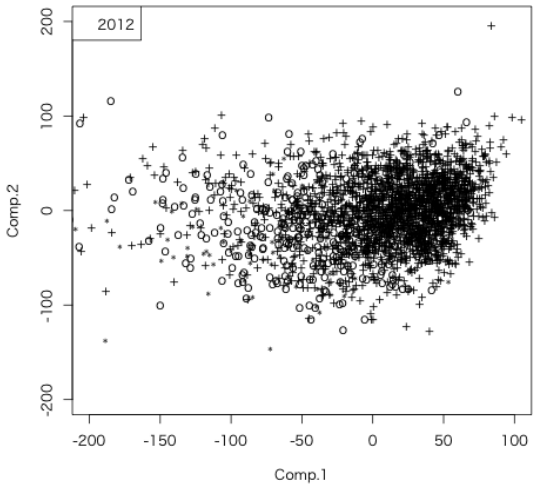

図 1　2012 年度血液検査プロット結果

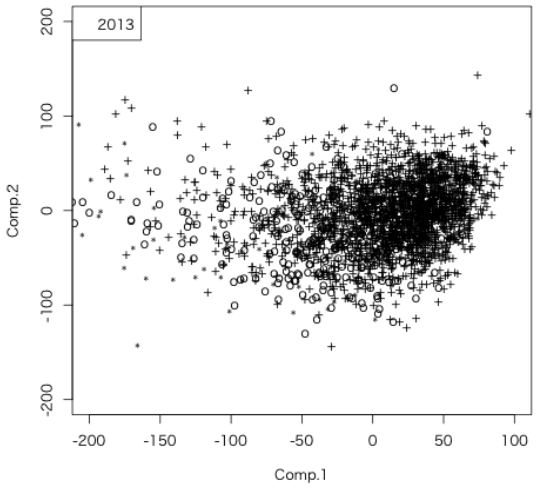

図 2　2013 年度血液検査プロット結果

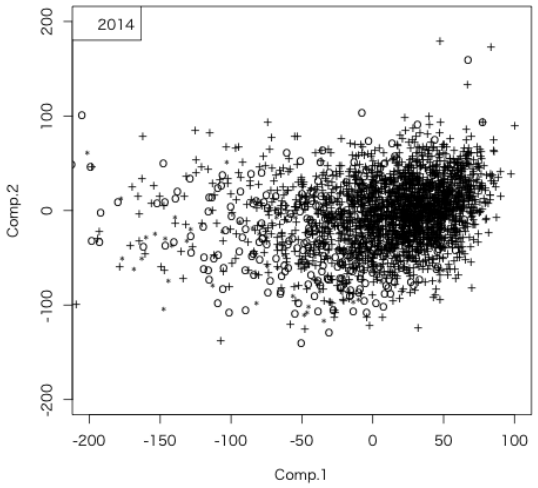

図 3　2014 年度血液検査プロット結果

6. まとめ

本研究では, 血液検査の多次元データを PCA を用いることによって, 二次元グラフ上にプロットし保健師指導効果の可視化を行った.

謝辞　本研究は JSPS 科研費 25750095 の助成を受けたものです.

参考文献

[1] 金森 敬文, 竹之内 高志, 村田 昇, “R で学ぶデータサイエンス”, 共立出版株式会社,pp.36–48, 2009.
[2] 金 明哲, “R によるデータサイエンス”, 森北出版株式会社,pp.66–77, 2007.