データマイニングを用いた最適プランの提案

2019SS052 南谷理穂
指導教員：佐々木美裕

1　はじめに

現代社会では物事の機械化が進み, それは実際的な作業だけでなく知能にも進出していることは周知の事実であり, AI の名は今や広く普及している. 昨今の機械学習研究は, 膨大なデータを基になされる予測が新たな視点をもたらすという意味からも期待されている.

本研究ではその中でも特にデータマイニングを取り上げ, 意思決定に係るプロセスを扱う. データマイニングについて, 藤野 [1] は『大量にあるデータから何かを見つけて, 現実社会において, あるいは未来社会において, 何かに役立てることをと考えるのがデータマイニングといえるでしょう』と述べている.

2　問題およびモデルの説明

都道府県に着目し, 類似性を発見すると共に各個人が持つ希望がどの都道府県と合致するのかを提案する. その方法として, 階層型クラスタリング・非階層型クラスタリング・実データと希望値の差の 2 乗和の計算を試して結果を比較しながら考察する.

都道府県データランキング [2] では, 各都道府県における膨大な量のデータを公開している. ここで公開しているデータのうち 78 項目を取り出して系統を統合し, データを作成する. この 78 項目は, 公開されているデータの中からクラスタリングなどの分析を行う際に系統の偏りがなるべく小さくなるようにまんべんなく選択したものである. 本研究でデータとして用いた項目を系統と共に表 1 に示す.

表 1　データとして用いた項目

人口系	地理・気象系	事故・交通系	観光・スポーツ系	健康・生活系
1. 人口密度	10. 年平均気温	20. 交通事故発生件数/10 万人	24. ホテル数	30. 持ち家比率
2. 出生率	11. 年平均湿度	21. 世帯あたり台数 (車)	25. 温泉地数	31. 1 住宅あたり宅地面積
3. 婚姻率	12. 年間降水日数	22. 1 人あたり台数 (車)	26. 博物・美術館数	32. 病院数/10 万人
4. 離婚率	13. 年間日照時間	23. 駅数 /100km^2	27. 水族館・動物園・植物園数	33. 1 人あたり医療費総額
5. 単独世帯率	14. 年間快晴日数		28. 県対抗駅伝男子平均順位	34. 喫煙率
6. 核家族世帯率	15. 年間雪日数		29. 県対抗駅伝女子平均順位	35. 美顔偏差値
7. 三世代世帯率	16. 可住地面積			36. 1 人あたりのビール消費量
8. 平均寿命 (男)	17. 森林面積			37. 育児をしている女性の割合
9. 平均寿命 (女)	18. 田面積			38. そのうちの有業率
	19. 宅地面積			39. 60 歳以上の有業率
社会系	**教育系**	**商業・通信・資源系**	**政治・行政系**	**工・農・水産業系**
40. 事業所数	49. 幼稚園児数/人口	59. 上位 10 チェーンコンビニ数/1 万人	64. 1 票の格差 (衆議院)	74. 製造業事業所数
41. 従業員数/事業所数	50. 保育数/園数	60. インターネット利用率	65. 歳入額 (100 万円)	75. 出荷額 (億円)
42. 最低賃金	51. 小学生数/人口	61. 年間通販利用額/人口	66. 歳出額 (100 万円)	76. 総農家数
43. 民営賃貸住宅の家賃	52. 児童数/小学校数	62. 1 人 1 日あたりごみ排出量	67. 地方議会合計女性議員比率	77. 総農家比率
44. 県民所得/人口	53. 高校数	63. ごみのリサイクル率	68. 第 24 回参議院議員選挙投票率 (18-19 歳)	78. 漁港数
45. 生活保護受給率 (全年代)	54. 生徒数/人口		69. 第 24 回参議院議員選挙投票率 (全年齢)	
46. 空き家率	55. 大学数		70. ふるさと納税寄付受総額 (千円)	
47. 検挙率	56. 短大数		71. 住民税減税額 (千円)	
48. 自殺率	57. 小中高暴力行為発生件数/千人		72. 1 人あたり寄付受額	
	58. 中学校不登校人数/千人		73. 1 人あたり減税額	

3　計算結果

3.1　階層型クラスタリング

最短距離法, 最長距離法, ウォード法の 3 種類の手法を用いて階層型クラスタリングを行う. それぞれ, 2 つのクラスタ間で 1 番近いデータ同士の距離をクラスタ間の距離として採用する方法, 2 つのクラスタ間で 1 番遠いデータ同士の距離をクラスタ間の距離として採用する方法, クラスタの各値からその質量中心 (重心) までの距離を最小化する方法である. 表 1 で示した項目すべてのデータを用いて実行した. ウォード法がこの中で最も詳細な階層分けを実現できたので, その実行結果をデンドログラムを用いて可視化したものを図 1 に示す.

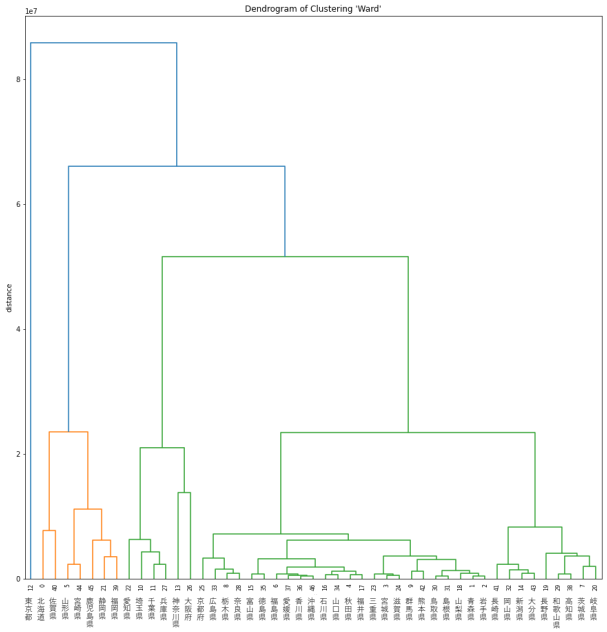

図 1　ウォード法クラスタリング実行結果

図 1 から, 東京都が他の要素と大別されていることがわかる. その他の階層では, 少しばらつきはあるものの神奈川県と大阪府が同一の階層にあることや, 千葉県・兵庫県・埼玉県・愛知県が近いところに位置していることから, 各都道府県の経済力が大きく寄与していると考えられる. 最短距離法・最長距離法でも東京都が大別される結果となったので, 東京都は特別な特徴を持つことがわかる.

3.2　非階層型クラスタリング

クラスタリングの手法については藤野 [1] のプログラムを参考に k-平均法を用いる. この手法ではクラスタ数を設定する必要があるので, 適切なクラスタ数をエルボー法を用いて求める. エルボー法とは, 各データが所属するクラスタ中心からの距離の 2 乗和 (誤差指標) に関して, クラスタ数を変化させながら誤差指標の変化を見る方法である. 誤差指標の値が小さいほど正確なクラスタリングが行われているといえる. 表 1 の項目すべてのデータを用いたエルボー法の計算結果を図 2 に示す. 横軸がクラスタ数で縦軸が誤差指標である. この結果より, クラスタ数が 8 以降の区間で誤差指標の値が安定したので, クラスタ数を 8 に設定する.

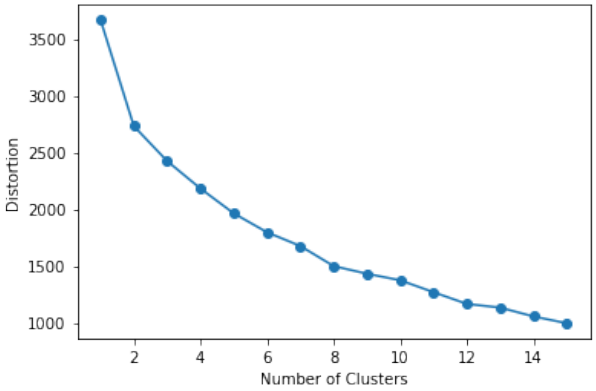

図 2　エルボー法計算結果

次に, 78 項目のうち任意の項目を選択してクラスタリングを行い, 入力する希望値に近い都道府県を求める問題について説明する. 例として, 表 1 の健康・生活系より 10 項目を選択し, 健康面に特化したクラスタリングを行う. 作成したプログラムでは, 選択した項目でのエルボー法実行や各項目におけるユーザの希望値入力が含まれている. ここでユーザの希望値とは, 各項目に対する重要度を 0 から 1 の範囲で正規化したものである. 希望値は項目 34. 喫煙率のみ 0 とし, その他 9 項目では 1 とした. つまり, 最も健康面を重視した条件である. また, 実行結果を図示するために正規化したデータに対して主成分分析を行い, 横軸を第 1 主成分, 縦軸を第 2 主成分としてプロット図を作成した. 主成分分析の寄与率は第 1 主成分, 第 2 主成分の順に 0.2749093, 0.21853392 となった. 累積寄与率は 0.49344322 である. 図 3 に 10 項目選択して実行した非階層型クラスタリングの結果を示す.

このとき希望値が属するクラスタは 2(青の丸印) であった. このことから, 健康面を重視する人にはクラスタ 2 に含まれる 8 つの要素である秋田県・山形県・新潟県・富山県・石川県・福井県・山梨県・島根県が適しているといえるだろう. またクラスタ 2 が最も健康的な要素が集まっているクラスタともいえる.

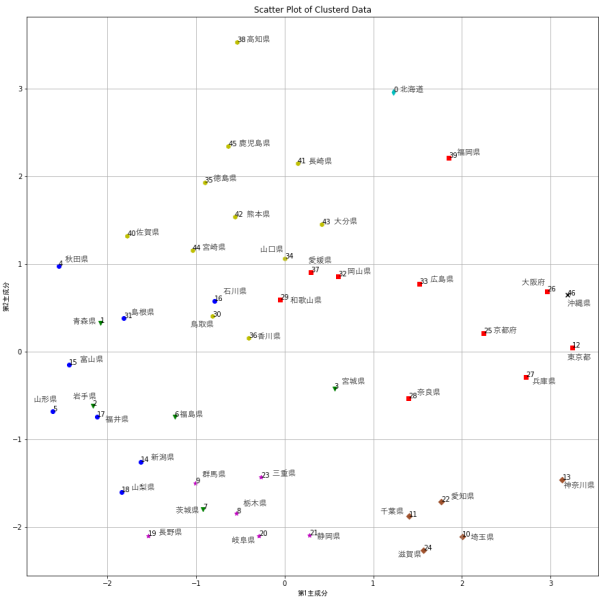

図 3　生活・健康系の項目を用いた非階層型クラスタリング実行結果

3.3　差の 2 乗和の計算による適切な都道府県の提案

実データと希望値の差の 2 乗和を求めるプログラムを作成した. 希望値は非階層クラスタリングと同様に 0 から 1 の範囲で正規化されたデータと正しく比較できるように設定するものとする. 項目選択した非階層型クラスタリングと同じ項目を用い, 希望値も同様に設定したときの実行結果を表 2 に示す.

表 2　差の 2 乗和計算結果

順位	都道府県名	差の 2 乗和
1	富山県	2.09834
2	島根県	2.12677
3	石川県	2.12766
4	宮崎県	2.27887
5	鳥取県	2.33168

先に示した非階層型クラスタリングの結果と照らし合わせて考察すると, 富山県・石川県・島根県はどちらの実行結果にも現れたことから, この 3 県は健康面を重視する人には適しているといえるだろう.

4　おわりに

3 種類の方法で都道府県のデータを分析した. 階層型クラスタリングや, 全項目を使用する非階層型クラスタリングでは, 大まかな区別として 1 番目に各都道府県の経済力, 2 番目に地理的関係がテーマとして挙げられると感じた. これらは普段の生活からも感じられるような区別であったので, 妥当な結果が出せたのではないかと考えている. さらに項目を絞った非階層型クラスタリングや希望値との差を見るプログラムから, 各個人の趣向を取り入れて考察することが可能になった.

本研究では, 都道府県のさまざまなデータを集め, 都道府県の特徴をさまざまな項目から比較・分類することができた. また, 項目ごとの重要度を表す希望値を定め, 希望に合致する都道府県を提案することができた. 課題として, プログラムの有効性をより確かなものにするために, その他のインスタンスのデータを用いてプログラムを実行することが挙げられる.

参考文献

[1] 藤野巌. データマイニングと機械学習. オーム社, 2019.

[2] M. Higashide. 都道府県データランキング. `https://uub.jp/pdr`, 2022 年 9 月 7 日閲覧.