多次元正規分布を用いた気象データの分類

小田部　修斗　　三浦　孝夫
法政大学理工学部創生科学科

1. 前書き

気象情報を利用したデータマイニングは様々な分野で利用されている。しかしながら、気象データは多次元データであり、多次元データマイニングは容易ではない[1]。

本研究では、気象データを対象とし、多次元正規分布モデルの推定と分類の手法を検証する。要素間の相関を考慮し、多次元正規分布モデルを最尤推定法によって作成し、モデルごとの尤度を比較することでデータの分類を行う。有効性を評価するため、無相関性多次元分布を用いる。

２. 多次元正規分布の最尤推定

本研究では、最尤推定法を用いて気象データの標本データから多次元正規分布モデルを推定する[2]。$\mathbf{x}=x_1,x_2,\ldots,x_n$ が発生する確率を表す確率密度関数 $f(\mathbf{x})$ を考える。これをパラメータの関数と見たものを尤度関数と呼ぶ。この尤度関数 $L(\theta)=f(\mathbf{x}|\theta)$ を最大化するようなパラメータの値を推定値とする方法が最尤推定法である。

多次元正規分布の最尤推定では、対応するパラメータで偏微分し、その極値を算出することで推定する。

平均 μ、分散共分散行列 Σ の d 次元多次元正規分布の確率密度関数 f(x)は、次の式で表される。

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

あるデータ $x_1,x_2,\ldots x_n$ が独立に f(x)に従うとき、対数尤度関数 $\log L(\mu,\Sigma)$を考え

$$logL(\mu,\Sigma) = \sum_{i=1}^{n} logf(x_i)$$

を各パラメータで偏微分し、極値を求めると。

$$\mu = \frac{1}{n}\sum_{i=1}^{n}(x_i)$$

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^T(x_i-\mu)$$

となり、多次元正規分布の最尤推定値は、データ集合の平均と分散共分散行列値と等しくなる。

3 実験

3-1 実験手順

本研究では、気象庁から提供されている、“過去の気象データ”を使用する。1987 年から 2016 年までの 30 年分のデータのうち、全国 8 地方の代表都市における 11 個の要素についてのデータを使用する。

解析ツール“R”を用いて、1992 年から 2016 年までの 25 年分のデータを学習データとして、8 都市の月毎、計 96 個の多次元正規分布モデルを最尤推定法で推定する。

比較対象(ベースライン)として、要素間に相関がないと仮定し、分散共分散行列の非対角要素を 0 に置き換えた多次元正規分布モデルを作成する。

テストデータとして、1987 年から 1991 年までのデータを使用し、データセットを作成する。

各テストデータにおいて、全てのモデルの尤度を計算し、本来分類されるべきモデルが尤度の大きい順で上位 3 位までの中に入っている場合を「正解」とし、正解率を比較し分類精度を評価する。

3-2 実験結果

表1に相関を考慮に入れた場合、相関を考慮に入れない場合のそれぞれの 1 位から 3 位までの正解数と割合を示す。

表 1　正解数と割合

	相関無考慮		相関考慮	
順位	正解数	割合	正解数	割合
1	60	31.3%	75	39.1%
2	25	13.0%	51	26.6%
3	23	12.0%	21	10.9%
計	108	56.3%	147	76.6%

相関を考慮しない場合の正解率は 56.3%(108/192)であるのに対し、相関を考慮した場合には 76.6%(147/192)という正解率となり、20.3%の向上が見られる。

3-3 考察

地点別の正解率を見ると、札幌の気温や松山の降水量などのように、1 つの要素に他と明らかに異なる特徴があった場合には相関を考慮しない場合でも正解率が高くなっている。対して、東京や広島のように、個別の要素には特徴が見られない地点では、相関を考慮するかどうかで正解率に大きく差が出ている。

これは、データを 1 つ 1 つの数値の集まりではなく、複数の数値の組み合わせとして見ることで更なる情報を引き出せるということを示している。

4. 結論

データの要素間の相関を考慮した多次元正規分布モデル作成と分類を行った。相関を考慮しない場合に比べ正解率が 20.3%向上し、分類精度が高くなった。

参考文献

[1] Han, J. et al.: Data Mining - Concepts and Techniques (3rd), Morgan Kaufmann, 2011
[2]平岡和幸・堀玄『プログラミングのための確率統計』オーム社　2009