MTシステムのカーネル化とその応用

リスク解析戦略研究センター　製品・サービスの質保証・信頼性研究グループ
特任研究員　　佐野　夏樹

1　はじめに

品質管理の分野において，良品はある程度の均一な集団を形成しても，不良品には多様な現象が混在し一つの意味のある母集団をなしているとは考えにくい．このような非対称な群（ラベル）から形成されるデータに対する判別問題を非対称判別問題と呼び，代表的な手法として，マハラノビス・タグチ (MT) システム (田口 (1995)) が知られている．

一方で機械学習の分野では最近，サポートベクトルマシンやサポートベクトル回帰などのカーネル手法が，提案され，様々な分野に適用されている．

本研究の目的は，MT システムのカーネル化を行い，MT システムよりも，柔軟な判別曲線を構築し，分類精度を向上させることである．また，実データとして，ワインの品質データを取り上げ，その有効性を検証する．

2　MTシステム

MT システムとは，単位空間と呼ばれる正常品の集合を定め，単位空間の標本平均と標本分散共分散行列を推定し，それをもとに，マハラノビス距離によって，個々のサンプルが，単位空間に属するか否かを判定する方法である．基本的な手続きは以下の通りである．

1. 各サンプルの標準化 $u_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, i = 1, \ldots, n, j = 1, \ldots, p$ を行う．ここで，$\bar{x}_j$ は標本平均，s_j は標本標準偏差である．
2. 各サンプルのマハラノビス距離 $D_i^2 = \boldsymbol{u}_i R^{-1} \boldsymbol{u}_i, i = 1, \ldots, n$ を計算する．ここで R は標本相関係数 $r_{ij} = \frac{1}{n} \sum_{k=1}^{n} u_{ki} u_{kj}, i, j = 1, \ldots, p$ を要素とする相関係数行列である．
3. D_i^2 が閾値，例えば $\chi^2(p, \alpha)$ を越えているかによって，単位空間に属するか否かを判定する．

3　カーネルMTシステム

MT システムは，元の入力空間において，マハラノビス距離により判別を行ため，暗に，単位空間が正規分布に従うことを仮定している．従って単位空間が正規分布に従わない場合には，適切に判別することが難しくなる．提案手法では，MT システムのカーネル化，すなわち，高次元もしくは，無限空間においてマハラノビス距離を計算することで，入力空間の単位空間がより，複雑な分布をする場合にも，適切な判別を行うことができる．高次元空間におけるマハラノビス距離として，Haasdonk, B. and Pekalska, E. (2008) による

$$d_{\mathcal{H}}^2 = \frac{1}{\sigma^2} (\tilde{k}_{xx} - \tilde{\mathbf{k}}_x^T \tilde{\boldsymbol{K}}_{\mathrm{reg}}^{-1} \tilde{\mathbf{k}}_x) \tag{3.1}$$

を用いる．ここで，$\tilde{\boldsymbol{K}}_{\mathrm{reg}} = \tilde{\boldsymbol{K}} + \alpha \boldsymbol{I}_n, \alpha := n\sigma^2$ は，正則化された中心化カーネル行列であり，$\tilde{\boldsymbol{K}} = \boldsymbol{HKH}, \boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ である．$\boldsymbol{K}$ はカーネル行列，$\boldsymbol{I}_n$ は恒等行列，$\mathbf{1}_n = (1, 1, \ldots, 1)^T$ であり，n は単位空間のサンプル数を示す．$\tilde{\boldsymbol{k}}_x = \boldsymbol{k}_x - \frac{1}{n}\boldsymbol{K}\mathbf{1}_n$ は中心化された $\boldsymbol{k}_x = (k(\boldsymbol{x}_1, \boldsymbol{x}), k(\boldsymbol{x}_2, \boldsymbol{x}), k(\boldsymbol{x}_n, \boldsymbol{x}))$ を示し，$\tilde{k}_{xx} = k_{xx} - \frac{2}{n}\mathbf{1}_n^T\boldsymbol{k}_x + \frac{1}{n^2}\mathbf{1}_n^T\boldsymbol{K}\mathbf{1}_n$ は中心化された $k_{xx} = k(\boldsymbol{x}, \boldsymbol{x})$ を表す．σ^2 は正則化パラメータである．

カーネル MT システムの手続きは以下の通りである．

1. 正則化パラメータ，カーネル，及びカーネルパラメータを設定する．
2. 単位空間のサンプルに対して，(3.1) 式に基づいて，高次元空間のマハラノビス距離を計算する．
3. 計算されたマハラノビス距離の 99% 点を閾値に設定する．
4. 未知のサンプルに対して，高次元空間のマハラノビス距離を計算し，閾値を越えていたら，異常サンプルと判定する．

4　ワインの品質問題への適用

本節では，ワインの品質評価データに対して MT システム, カーネル MT システムを適用した結果を示す．使用したデータは UCI Machine Learning Repository の Wine Quality Data Set の白ワインデータである．11 項目のワインの化学的成分を特徴量とし，ワインエキスパートによる 10 段階の品質評価結果を分類ラベルとするデータセットである．サンプル数は 4898 であるが，3898 サンプルを学習データに，1000 サンプルをテストデータに分割する．10 段階の品質評価のうち，データセットにはランク 3 からランク 9 までのサンプルが存在し，ランクが高い程，高品質のワインである．そこで，ランク 8 とランク 9 の高品質ワインを単位空間 (ラベル +1), それ以外を非単位空間 (ラベル −1) に再ラベル付けを行った．学習データを用いて，MT システム (MTS) とカーネル MT システム (KMTS) における閾値などのパラメータを決定し，テストデータによって評価した結果を表 1 と表 2 に示す．カーネル MT システムのオッズ比 (759.92) は，MT システムのオッズ比 (22.88) よりも，約 33 倍，高い数値を示していることがわかる．これは，単位空間の分布，すなわち，高品質のワインの分布が正規分布には従っていないため, MT システムよりも，カーネル MT システムの方が高い判別性能を示したと考えられる．

表 1：MTS の予測結果とオッズ比

	MTS の予測結果	
実際のラベル	ラベル −1	ラベル +1
ラベル −1	199	758
ラベル +1	0	43

オッズ比：22.88

表 2：KMTS の予測結果とオッズ比

	KMTS の予測結果	
実際のラベル	ラベル −1	ラベル +1
ラベル −1	957	0
ラベル +1	31	12

オッズ比：759.92

参考文献

田口玄一 (1995). パターン認識のための品質工学 (3)，品質工学，**3**, 2-5.

Haasdonk, B. and Pekalska, E. (2008). Classification with Kernel Mahalanobis Distance Classifiers, *Advances in Data Analysis, Data Handling and Business Intelligence*, 351-361.

http://archive.ics.uci.edu/ml/datasets/Wine+Quality