4L-3

量質混在のデータに適用可能な一般化されたミンコースキー距離

矢口博之，岸　透，河村知之，戸塚伸吾，市野　学
東京電機大学

１．はじめに

パターン認識やデータ解析の分野において距離関数は主要な役割を果たしてきた．

判別分析の一つである最近隣決定法では２つのサンプル群の代表的なサンプルを記憶しておき，距離計算によって帰属すべき群を決定する．クラスター分析は，個体間の距離にもとづいて，似たようなサンプルを集め１つのグループ（クラスター）を作り，全体をいくつかのクラスターに分割することによってデータの構造を把握するための手法である．

連続量で表されるデータに対する距離としてミンコースキー距離やマハラノビス距離が，また二値データに対する距離としてはハミングの距離などが知られている．一方，我々は量的特徴と質的特徴が混在する形で与えられるデータに対して一般化されたミンコースキー距離を定義した[1]．

この報告では，一般化ミンコースキー距離を修正し，区間を値としてとるデータに対し，より自然な性質を持つ距離の定義を示す．

２．カルテジアン空間上での一般化ミンコースキー距離の定義

我々は，先の報告[1]で量的特徴と質的特徴の混在を許す特徴空間の上で，一般化されたミンコースキー距離を定義した．

各サンプル$\mathbb{X}$は，ｄ個の特徴の組で記述されているとする．全サンプルからなる集合を特徴空間とよび，$U^{(d)}$であらわすことにする．ここでいう特徴とは，

(1)連続値をとる量的な特徴
（身長，体重，血圧など）
(2)離散値をとる量的な特徴
（年齢，人数など）
(3)順序の入った質的な特徴
（学歴，年号など）
(4)名義的な特徴
（性別，血液型など）

である．

特徴空間上のサンプルＡとＢに対して，ＡとＢを合わせてより抽象度の高い概念を生成する働きを持つ演算であるジョイン⊞と，ＡとＢに共通な概念を抜き出す働きを持つ演算，ミート⊠を定義する．２次元平面上のジョインとミートの例を図－１と図－２に示す．

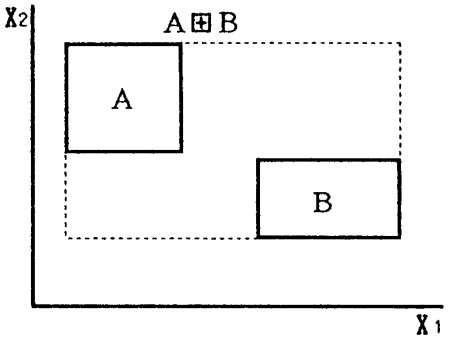


図-１ カルテジアン・ジョインの例

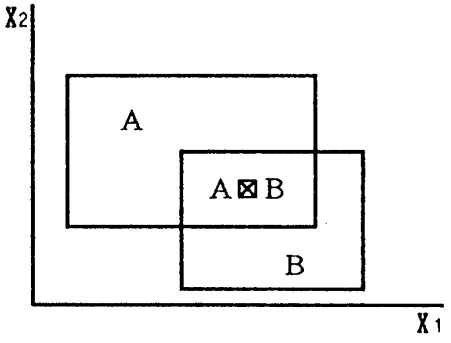


図-２ カルテジアン・ミートの例

特徴空間$U^{(d)}$と，この上に定義される演算⊞（ジョイン）と演算⊠（ミート）の組（$U^{(d)}$,⊞,⊠）をカルテジアン空間とよんでいる。

Generalized Minkowski Metrics for Mixed Features
Hiroyuki YAGUCHI,Toru KISHI,Tomoyuki KAWAMURA,Shingo TOZUKA,Manabu ICHINO
Tokyo Denki University

カルテジアン空間上のサンプルAとBに対して一般化されたミンコースキー距離 $d_P(A, B)$ は，

$$d_p(A,B)=\left[\sum_{k=1}^{d}\left\{C_k\frac{\rho(A_k,B_k)}{|U_k|}\right\}^p\right]^{1/p}$$

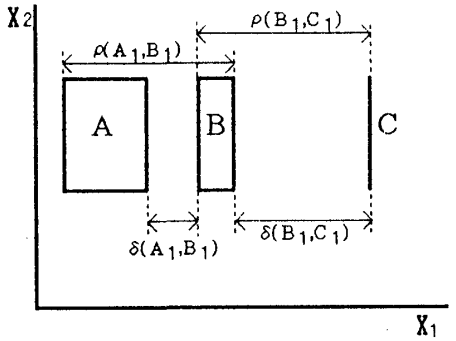
と定義される．ただし，

$$C_k>0,\ k=1;2,\cdots,d,\ \sum_{K=1}^{d}C_k=1$$

$$\rho(A_k,B_k)=|A_k\boxplus B_k|-|A_k\boxtimes B_k|$$

$|X_k|$ は連続値をとる量的特徴のときは X_k の長さ，離散値をとる量的特徴，順序の入った質的特徴のときと名義的特徴のときは X_k に含まれる可能な値の数であるとする．

しかし我々の定義では2つのサンプル間の最も遠い部分に注目した距離となっており，2つのサンプルの間の部分は共通部分がない限り距離に反映されない．たとえば図－3におけるサンプルA，B間とB，C間の距離は等しくなってしまう．従ってサンプルの間の部分を何等かの形で距離に反映させるには距離関数の形を変える必要がある．



図－3　3つのサンプル間の距離

3．一般化ミンコースキー距離の修正[2]

ここで特徴 X_k におけるサンプルAとサンプルBの間の部分 δ は

$$\delta(A_k,B_k)=|A_k\boxplus B_k|+|A_k\boxtimes B_k|-|A_k|-|B_k|$$

で表すことができる．そこで ρ と δ を加え正規化のために2で割ったものを新たに ϕ として定義する．

$$\phi(A_k,B_k)=\frac{\rho(A_k,B_k)+\delta(A_k,B_k)}{2}=|A_k\boxplus B_k|-\frac{|A_k|+|B_k|}{2}$$

この $\phi(A_k, B_k)$ を一般化されたミンコースキー距離 $d_P(A, B)$ の $\rho(A_k, B_k)$ の代わりに用いたものを修正された一般化ミンコースキー距離 $D_P(A, B)$ として次のように定義する．

$$D_p(A,B)=\left[\sum_{k=1}^{d}\left\{C_k\frac{\phi(A_k,B_k)}{|U_k|}\right\}^p\right]^{1/p}$$

距離関数をこのように変更しても距離の3公理を満足することは文献[1]の証明法に従って示すことができる．

4．まとめ

量的特徴と質的特徴の混在する形で与えられるデータに対して適用可能な一般化ミンコースキー距離を修正し，区間を値としてとるデータに対して自然な性質をもつ距離を提案した．この距離を用いることでデータ解析やパターン認識の分野の各種の方法を一般化し，より一般性の高い概念を扱うことができると考えている．

文　献

[1] 市野，矢口：”量質混在の記述を許す一般化されたミンコースキー距離” 信学論（A），J72-A,2,pp.398-405 1989年2月

[2] 小野：”量質混在データに適用可能な一般的クラスタリング法の研究”，東京電機大学大学院理工学研究科システム工学専攻修士論文，1989年2月