

はじめてのパターン認識

第5章 k最近傍法

大内克之

最近傍法

入力データを、最も距離の近い学習データ(鑄型ともいう)が属するクラスに分類する方法を**最近傍法**という。

入力データと距離が最も近いk個の鑄型を選び、所属する鑄型が最も多いクラスに分類する方法を**k最近傍法(kNN法)**という。

鑄型がもつ支配領域を**ボロノイ領域**という。

kNN法の計算量低減法

投票型k最近傍法では、全ての鑄型とのユークリッド距離の二乗を求める必要がある。

この計算量を低減する方法には、下の3つがある。

誤り削除型kNN

圧縮型kNN

分岐限定法

近似最近傍探索

誤り削除型kNN

クラスAのデータがクラスBの領域に存在している場合、正しい領域にあるデータも誤ってしまう。

そこで、誤って識別されたデータを全て削除する。この学習データセットを誤り削除集合という。

これを用いて行うkNN法を誤り削除型kNNという。

圧縮型kNN

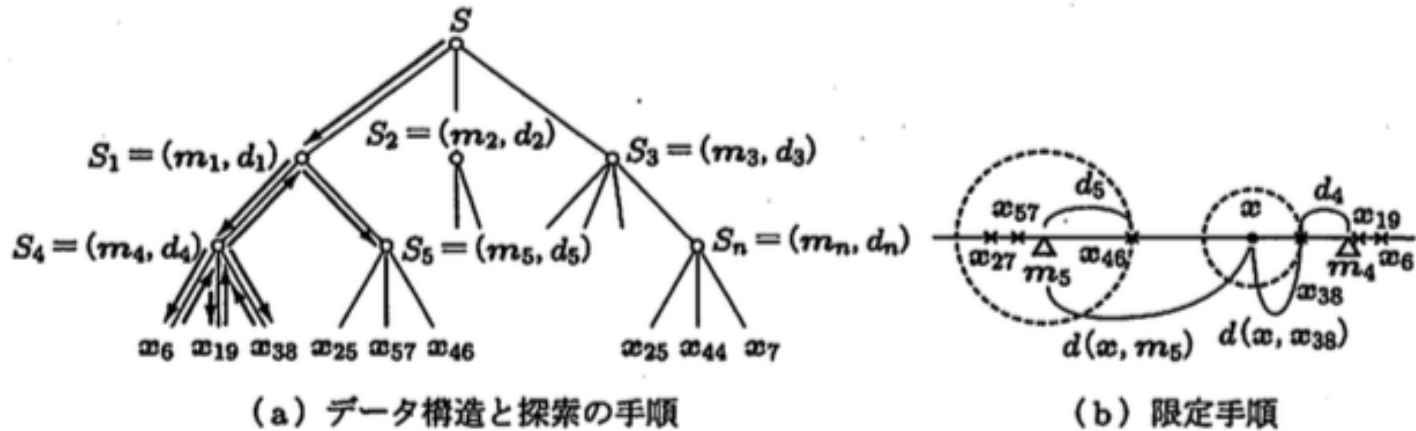
識別に寄与しないデータを学習データから削除するkNN法を
圧縮型kNNという。

分岐限定法

学習データをクラスタリング法などで重ならないクラスに分割し、木構造に組織化する。

各ノード下の学習データがそのノードに属する集合となる。このような木構造を作る方法を**分岐法**という。

分岐限定法



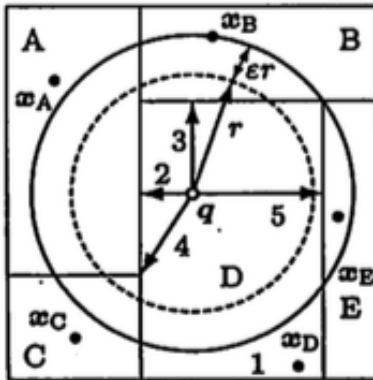
入力データ x が与えられると、まず図(a)のように S_1 から S_4 まで辿り、集合内の学習データとの距離を求める。

最も近いものが x_{38} だとすると、

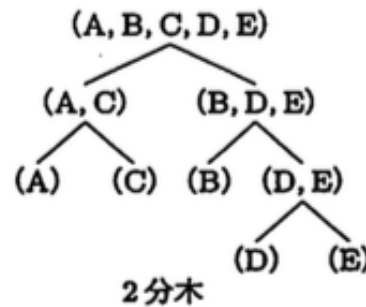
$$d(x, m_5) > d(x, x_{38}) + d_5$$

となるとき、 S_5 の計算を省く。これを**限定法**という。

近似最近傍探索



(a) 学習データの所属領域



(b) 最良優先探索

最近傍探索は、次元が大きくなると遅くなる。

そこで、近似解を高速で求める**近似最近某探索**に関する研究が行われてきた。

近似最近傍探索には、最良優良探索などがある。