

増減する点オブジェクトに対する 近接クラス集合の集計アルゴリズム

田中直幸

m073263@hiroshima-u.ac.jp

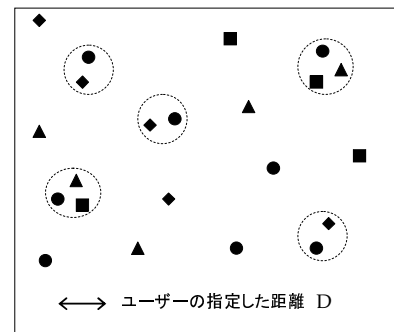
森本康彦

morimoto@mis.hiroshima-u.ac.jp

(広島大学大学院工学研究科)

1 はじめに

近年、空間データを含むデータベースから空間的文脈で頻出するパターンを発見する空間データマイニングの研究が進められている。そこで、2次元空間上に分布している点オブジェクトの頻出パターンに対し、点オブジェクトが追加、削除される状況を想定し、頻度指標であるサポート数の計算手法を提案する。これにより点オブジェクトの分布が変化した場合、毎回初めからアプリオリ生成を再実行する操作を回避でき、計算量を軽減できる。



2 近接クラス集合

本論文では図1のように2次元平面上に分布する点を対象とする。それぞれの点は座標と属性の値を持ち、それらを点オブジェクトと呼ぶことにする。例えば、携帯電話などの位置情報サービスでは、ある地点で利用したサービスの種類などが属性に該当する。各点オブジェクトは属性値によってクラス分けされる。図1では、●、▲、■、◆がそれぞれの点のクラスを表している。

「近接クラス集合」とは、その集合内のオブジェクトが互いにユーザーの指定する距離 D 以内に存在するものをいい、近接クラス集合のインスタンス数をサポート数と呼ぶ。近接クラス集合のサポートを

$$(\{c_1, c_2, \dots, c_k\}, \text{サポート数})$$

と記述する。ここで c_1, c_2, \dots, c_k はこの近接クラス集合の要素を表す。図1ではそれぞれの円に含まれるオブジェクトが距離 D 以内に分布し、●、▲、■のクラス集合が2箇所、●、◆のクラス集合が3箇所距離 D 以内に近接しているため、これらの近接クラス集合のサポートは $(\{\bullet, \blacktriangle, \blacksquare\}, 2)$ 、 $(\{\bullet, \blacklozenge\}, 3)$ となる。近接クラス集合のうちユーザーの指定する最小サポート数

図1: 近接クラス集合

以上の集合を「頻出近接クラス集合」とし、この頻出集合を数え上げる。

3 頻出近接クラス集合数え上げ

頻出近接クラス集合を数え上げる手法 [2] はアプリオリアルゴリズム [1] を応用しており、手順の概要をアルゴリズム1に示す。

アルゴリズム1 頻出集合数え上げ

入力: 点オブジェクト集合

出力: 全頻出近接クラス集合

- (1) $k = 1$ と初期化
- (2) k 候補集合のサポート数カウント
- (3) 頻出 k 近接クラス集合を求める
- (4) 頻出 k 近接集合から $(k + 1)$ 候補集合を生成
- (5) $k + 1$ 候補集合が空集合のとき、終了
空でないとき、 $k = k + 1$ として (2) にもどる

ステップ (4) の頻出 k 近接クラス集合から $(k+1)$ 候補集合を生成する際には、ある点オブジェクトが、あるクラスのオブジェクトのどの点に最も近いかを頻繁に計算する必要があるがこの部分ではオブジェクトに対してあらかじめボロノイ図と呼ばれるインデックス構造を作っておくことで効率的になる。

4 増減する点オブジェクトに対するサポート計算

上で述べた頻出近接クラス集合の数上げは問題の入力として点オブジェクト分布がすべてがはじめから与えられているが現実のエリアマーケティングや位置情報サービスではオブジェクトの追加、削除があり、その度に頻出パターンを再計算する必要がある。そこで現時点で存在する点オブジェクトに対する頻出近接クラス集合はすでに求められていると仮定し、そこに1つの点オブジェクトが追加、あるいは削除されたときのサポートの再計算アルゴリズムを考える。

まず1つの点オブジェクトが加えられる状況を想定する。新たに点オブジェクトが追加されると1近接クラス集合でいずれかの要素のサポート数を1だけプラスする。そこで、もし頻出1クラス集合が更新されていればその更新されたクラス集合をもとに2クラスの候補集合をつくる。更新されなければサポートの計算は終了する。つまり頻出 k クラス集合が更新されている間はアプリアリ生成を続けていく。ただし、追加した点オブジェクトと距離 D 以内にあるクラス集合をあらかじめ特定しておくことで $(k+1)$ 候補集合を生成する過程での計算を節約できる。以下にそのアルゴリズムを示す。

アルゴリズム 2 オブジェクト追加のサポート計算

入力：1近接クラス, 全頻出近接クラス
1点オブジェクト

出力：全頻出近接クラス

- (1) 加えられた点クラスの1近接クラス集合のサポートをプラス1
- (2) 頻出1近接クラス集合が変化しない場合、終了。変更された場合は $k = 1$ として (3) を実行
- (3) 更新されたクラス集合で、かつ追加点との距離が D 以内のものを対象にして $(k+1)$ 候補集合生成。

- (4) $(k+1)$ 候補集合のサポートをカウントして、頻出 $(k+1)$ 近接集合をもとめる。
- (5) 頻出 $(k+1)$ 近接集合が変化した場合、(3) にもどる

このアルゴリズムでは入力として全頻出クラス集合に加えて1近接クラス集合も与えているがこれは、1近接クラスのサポートの集計を省くためである。

最小サポートが3で、1近接クラス集合が $(\{\bullet\},6),(\{\blacktriangle\},5),(\{\blacksquare\},3),(\{\blacklozenge\},1)$ の例を考える。頻出1近接クラス集合は $(\{\bullet\},6),(\{\blacktriangle\},5),(\{\blacksquare\},3)$ である。この状態で \blacklozenge クラスのオブジェクトが与えられた場合、 \blacklozenge クラスのサポート数を $(\{\blacklozenge\},1) \rightarrow (\{\blacklozenge\},2)$ のようにプラス1するが頻出1近接クラス集合自体に変化はないのでサポートの計算を終了する。次に同じく \blacklozenge クラスのオブジェクトが追加されると頻出集合に $(\{\blacklozenge\},3)$ が加わり頻出集合が更新されたため、この $(\{\blacklozenge\},3)$ と追加点オブジェクトからの距離 D 以内のクラスを対象にして次の候補集合を生成する。また \blacktriangle が追加されたときには、頻出集合の要素数に変化はないが \blacktriangle のサポート数が $(\{\blacktriangle\},5) \rightarrow (\{\blacktriangle\},6)$ と変化しているのでこの場合にもアプリアリ生成の過程を続けていくことになる。

ここまでは点オブジェクトの追加のみを考えてきたが削除の場合にも同様な手順を適用できる。まず1つのオブジェクトが削除された場合、1近接クラス集合の一つの要素のサポート数をマイナス1する。そこで頻出集合に変化があればアプリアリ生成を続けていく。

このアルゴリズムは1点オブジェクトの追加、削除に対してサポートを再計算するがこの手順をオブジェクト数が連続的に変化する場合に適用することで順次、サポートの計算が可能で頻出集合を数え上げることができる。

参考文献

- [1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of VLDB Conference*, pages 487–499, 1994.
- [2] Y. Morimoto. Mining frequent neighboring class sets in spatial databases. In *Proceedings of ACM-SIGKDD Conference*, pages 353–358, 2001.