1T－7

データ空間の分割による数値データに関する相関ルールの抽出の効率化

猪鹿倉知広
東京大学大学院　工学系研究科
電子工学専攻

浜田　喬
学術情報センター

1　はじめに

近年さまざまな分野においてデータベースの数とそこに保存されているデータの数が急速に増加している。また、プロセッサ性能の飛躍的な向上と半導体メモリ技術の進歩により、膨大な量のデータをより積極的に解析することが可能になった。そこでデータベースに蓄積された膨大なデータから自動的に有用な情報を発掘する、データマイニングの必要性が高まっている。

データマイニングとはデータの中に隠されている情報を発掘することであり、対象となるデータベース、データの種類や得ようとする情報によって用いるべき手法は大きくことなる。従ってそこに必要な技術はデータベース技術、データ視覚化技術、統計科学的手法、人工知能の機械学習技術、推論技術など極めて広範な分野に及ぶ。そのなかでもデータマイニングを代表する手法として相関ルールの発見があげられる。

相関ルールとは例えばコンビニエンスストアなどでの「パンを買う人は同時に牛乳を買っていくことが多い」、というようなルールのことである。このようなルールを選ぶ基準としてはパンと牛乳を同時に買っていく人がある程度の数いること。及び、パンを買っていく人のうち牛乳を買っていく人の割合が充分多いことの二つになる。

この相関ルールを年齢、預金残高等の数値データに適応することによって、例えば「年齢が３０歳から３５歳で年収が８００万円ぐらいの人は貯金額が１００万円ぐらいである」等のルールを導き出すことができる。このような情報は銀行、クレジット会社などの顧客情報をつかむ上で非常に役立つ可能性を持っている。

Optimization of mining association rules for numeric data by dividing data space
Igakura Tomohiro[1],Hamada Takashi[2]
[1]University of Tokyo
[2]National Center for Science Information Systems

2　相関ルール

相関ルールの定義を行う。ここでアイテムの集合を$I = \{i_1, i_2, ..., i_m\}$、トランザクションデータベースを$D = \{t_1, t_2, ..., t_n\}(t_i \subseteq I)$とする。各要素トランザクション$t_i$はアイテム集合である。長さkのアイテム集合とは、k個のアイテムの組合せを指す。

それぞれコンビニエンスストアの例を使って説明すると、アイテムとは商品のことであり、トランザクションはレシート一枚一枚に相当する。

相関ルールは$X \Rightarrow Y$で表現される。ここで$X, Y \subset I$、XとYは同じアイテムを含まないとする。相関ルールは支持度および確信度の二つのパラメータを持ち、これらの値によって相関ルールの有意性を示す。支持度とは、D全体に対しXとYを共に持つトランザクションの割合、確信度とはXを含むトランザクションのうちYを含むトランザクションの割合によって与えられる。

つまり、Xの支持度を$support(X)$と表記すると$X \Rightarrow Y$の確信度は

$$\frac{support(X,Y)}{support(X)}$$

となる。

相関ルールを発見するとはユーザによって指定された最小支持度、と最小確信度を満足するすべてのルールを見い出すことに相当する。

2.1　数値データに関する相関ルール抽出の問題点

＜属性、値＞もしくは＜属性、値の範囲＞というように、数値データの値をクラスタに分け、それぞれのクラスタを二値データの時のアイテムとして使うことによって二値データの相関ルールと同様に求めることができる。

しかしこれには問題点がある。

- 一つ一つの数値の範囲が狭過ぎる場合、どの値も最小支持度を満たさない場合がある。
- 逆に範囲が広過ぎる場合、最小確信度を満たさなくなる場合がある。

このように相関があるデータであっても値の区切り方によって発見できなくなるということが考えられる。

上の問題を解決するための方法として小さいクラスタに分割し、それが接続したものについて相関ルールを抽出するというものが考えられる。

この方法の解決すべき問題点としては以下のようなものがある。

- ある属性を n 個に分割した場合属性ごとに実行時間が $O(n^2)$ になる。
- 最小支持度、最小確信度を満たす領域を大量にとり出してしまう。

3 求めるべきルールと分割法

どのようなルールを取り出すべきかというと

$$< x, a, b > - >< y, c, d >$$

(x が a から b の範囲にあれば y が c から d の範囲にあることが多いというルール)

$$< x, a, b > - >< y, d, e >$$

という二つのルールがある場合

$$< x, a, b > - >< y, c, e >$$

の支持度

$$\frac{support(< x, a, b >, < y, c, e >)}{support(< x, a, b >)}$$

から予想される支持度

$$\frac{support(< y, c, d >)}{support(< y, c, e >)} \frac{support(< x, a, b >, < y, c, e >)}{support(< x, a, b >)}$$

と実際の支持度がほぼ等しければ上の二つのルールと下の一つのルールの意味は同じであるといえるのでこれを分割する必要はないということである。つまり取り出すべきルールは下のルールであるということになる。

4 MDL を用いた分割法

MDL とは「与えられたデータを、モデル自身の記述も含めて含めてもっとも短くできるような確率モデルが最良のモデルである」と主張するものである。

考え方としては実際のデータがある確率密度関数によって作られるデータとして、データから元の確率密度関数を見つけるということである。相関ルールを求める際に問題となるのは y 軸なので、とりあえず y 軸のみを分割することを考える。

上の条件を満たすような分割として次のようなものを考える。元の確率密度関数を $f(X,Y) = q(Y)p(X|Y)$ とする。そして $p(X|Y)$ の等しくなるよう範囲を求める。このようにすればとなりあうクラスタ間において得られる相関ルールはことなるものになると考えられる。

このような分割も MDL を利用して求めることができる。

候補となる確率モデル集合として考えられるすべての分割の集合とし、それぞれのクラスタ内では $p(X|Y)$ が等しいとしこの中から、与えられたデータの符号長をもっとも短くする確率モデルを探し出すことによって得られる。

5 おわりに

本稿では相関ルールを数値データに適用しようとした際の問題点を述べ、それを解決するデータ空間の分割法を提案した。

今後の課題としては適切な分割法とは適応させるアプリケーションに大きく依存するので具体的なアプリケーションの特徴に合わせてどのように分割法を変えていくかを考えなければならない。

参考文献

[1] Agrawal,R., Srikant,R.: Fast Algorithms for Mining Association Rules, Proc. of VLDB,pp.487-499(1994)

[2] R. Srikant and R. Agrawal:Mining Quantitative Association Rules in Large relational Tables,SIGMOD Record,Vol.25,No.2,pp.1-12(1996)

[3] T. Fukuda, Y. Morimoto, T. Tokuyama:Mining Optimized
Association Rules for Numeric Attributes,Proceeding of ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems POD 1996

[4] 山西健司、韓太舜:MDL 入門:情報理論の立場から、人工知能学会誌、Vol.7,No.3,pp.427-434(1992)

[5] 韓太舜、小林欣吾:岩波講座応用数学 情報と符号化の数理、岩波書店 (1994)