ニューラルネットワークによるデータマイニング

5M－2

池田 智仁　福本 光　呉 旭　阿江 忠

広島大学 工学部

1　はじめに

近年、大量に蓄積された履歴データを解析することにより、その中に埋もれた法則や関係など有用な情報を抽出し、データの効果的な活用を図る研究として、データマイニング [1] が着目されている。

データマイニングで得られる情報の代表的なものとして、データ間の相関関係が挙げられる。このようなルールを抽出する場合、データベースを繰り返し検索する必要があり、また探索空間は 2^p(p:アトリビュート数) となるため、非常に負荷の高い処理となる。その要因を分析すると、

1) データベースから知識の最小表現を求める手数が一般には NP 完全である。
2) データベース内に互いに競合する無矛盾集合が存在し、その中から極大なものを求める問題も NP 完全である。

通常は 2) を包含した 1) の問題の議論をしているが、本稿では 2) の要因に着目し、問題の近似解法を容易にする方法を提案する。具体的にはホップフィールドニューラルネット [4] による近似解法が用いられる。

2　無矛盾集合に着目したデータマイニング

本稿では数値表形式のデータベースからの知識獲得を考える。データ構造を次のように定義する。

$$data(l) : (x_{1l}, x_{2l}, \cdots, x_{pl}) \tag{1}$$

本稿の目標は次の問題を解くこととする。

[問題 1] (1) で定義されるデータベースから、
$x_i \rightarrow x_j, x_i \rightarrow \bar{x_j}$のようなシングルクローズ形式のルールの集合の最小表現を抽出する。

この問題を解くために各リテラルの真偽を定義する。そのために、データベースのアトリビュートごとにあるしきい値を定め、それ以上であれば“真”、そうでなければ“偽”とし、データベースを2値化する。また、全ての(正の)リテラル $x_i(i = 1, 2, \cdots, p)$ をベクトル表現 $X = (x_1, x_2, \cdots, x_p)$ し、x_iの値を1(真)、0(偽) で表しその集合を Sとする。すると、[問題 1] は [問題 2] のように変換できる。

[問題 2] Sの中から真 (無矛盾) な集合を求め、その集合を表現する最小のルールの集合を求める。

Sに含まれる無矛盾な部分集合 S_kは数多く存在するが、その中で Xの重み (1 の数) を最大とする S_iは極大無矛盾部分集合である。ここで、[問題 2] は [問題 3] のように変換される。

[問題 3] S_kの中から極大無矛盾部分集合を求め、その集合を表現する最小のルールの集合を求める。

極大無矛盾部分集合を求める問題は、NP 完全問題である。本稿ではその近似解法として、データベースの相関関係をグラフ化し組合せ最適化問題に置き換えて、その問題をニューラルネット [3] を用いて解くことにより極大無矛盾部分集合を求める手法を提案する。ニューラルネットとしては、組合せ最適化問題に対し高速に近似解を得ることが知られているホップフィールドニューラルネットを用いる。

3　ホップフィールドネットへの写像

[問題 3] を解くため、リテラル間の相関関係を導入する。ここで [問題 3] は次の [問題 4] に変換される。

[問題 4] データベースより、リテラル間の相関関係を求めてから極大無矛盾部分集合を求め、その集合を表現する最小のルールの集合を求める。

リテラル間の相関関係 $a_{ij}(i \neq j)$ を以下のように定義する。

$$a_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} = 1)(x_{jk} = 1)}{\sum_{k=1}^{n}(x_{ik} = 1)} \tag{2}$$

これより相関マトリクスが得られる。

まず、相関関係が 2 値で表現されているものとして考える。ここで以下の 2 つを定義する。

- $a_{ij} = 1$ のとき、リテラル x_iが真ならば、リテラル x_j は真である。
- $a_{ij} = 0$ のとき、リテラル x_iが真ならば、リテラル x_j は偽である。

これは次式のように表現することができる。

$$(x_i \rightarrow x_j a_{ij} \vee \bar{x_j}\bar{a_{ij}}) = 1 \tag{3}$$

これより、すべてのリテラルのつくる状態は

$$U = \prod_{i,j \in E} (\bar{x_i} \vee x_j a_{ij} \vee \bar{x_j}\bar{a_{ij}}) \tag{4}$$

となる。この式 (4) はすべてのリテラルが式 (3) を満たすとき、$U = 1$ となる。

以上は論理表現であり、式 (4) は 1(真) か 0(偽) の値をとる。本稿では相関マトリクスの値に [0,1] の間の有理数を導入しているため、拡張が必要である。

Data Mining by Neural Network
Tomohito Ikeda, Hikaru Fukumoto, Wu Xu and Tadashi Ae
Fuculty of Engeneering, Hiroshima University

[問題5] 矛盾量最小の集合を求め、その集合を表現する最小のルールの集合を求める。

すなわち、相関マトリクスが有理数になったため、完全に無矛盾な極大部分集合は通常存在しなくなる。従って、制約条件を緩和して矛盾量最小なものを最適解として求めることにする。

まず、真であるリテラル数が最大である解を求める問題を、双対変換して、最小化問題として解く。

$$f_1 = \text{矛盾を表す項(無矛盾であれば0)} \tag{5}$$

$f_2 =$ 真であるリテラルの数が規定の数に達すれば0になる項 (6)

とし、次式の最小化を目的とする。

$$f = Af_1 + Bf_2 \tag{7}$$

但しAとBは重み付けのための定数である。式(4)を変形して

$$\begin{aligned}
\bar{U} &= \sum_{i,j\in E}(x_i(\bar{x_j} \vee \bar{a_{ij}})(x_j \vee a_{ij})) \\
&= \sum_{i,j\in E}(x_i\bar{x_j} \vee x_i\bar{a_{ij}})(x_j \vee a_{ij}) \\
&= \sum_{i,j\in E} x_i\bar{x_j}x_j \vee x_ix_ja_{ij} \vee x_ix_j\bar{a_{ij}} \vee x_i\bar{a_{ij}}a_{ij} \\
&= \sum_{i,j\in E} x_i\bar{x_j}a_{ij} \vee x_ix_j\bar{a_{ij}}
\end{aligned}$$

ここでx_i及びa_{ij}は、$[1,0]$の間の有理数をとるとし、これを数値化すると

$$f_2 = \sum(x_i(1-x_j)a_{ij} + x_ix_j(1-a_{ij})) \tag{8}$$

となる。これと式(6), 式(7)を併せて次式を得る。

$$f = A\sum(x_i(1-x_j)a_{ij} + x_ix_j(1-a_{ij})) + B\sum(1-x_i) \tag{9}$$

[問題5]を具体化すると、fを最小化するベクトルXを求めるということになる。

4 ニューラルネットモデル

生物の脳は、ニューロンを基本単位とし、そのニューロン間は、シナプスにより信号を伝達する。このニューロンを模倣した人工のニューラルネットワークを考える。(以下、ニューラルネットおよびニューロンはそれぞれ人工のニューラルネットワーク、ニューロンを指す。)

ニューロンは多入力一出力のしきい値素子で、他のニューロンからの入力x_iに重みT_{ij}を掛けたものの総和としきい値とを比較し、非線形的に出力するものである。

ニューラルネットには、ニューロン間で相互に情報を流すことができ、あるニューロンが伝えた相手のニューロンからも情報をフィードバックできる相互結合型と、ニューロン間の信号の流れが一方通行で、入力層へ入力された情報が、隣接する層のニューロンへ順次伝搬される多層型のモデルがある。本報告では、組合せ最適化問題の解法に向いている相互結合型のホップフィールドニューラルネットを用いる。

ホップフィールドニューラルネットは、以下の式(10)により求められるニューロンの内部活性値v_iをもとに、式(11)のシグモイド関数に従ってニューロンの出力x_iを更新する。ここでニューロンの出力はλの値により0と1の間の連続値をとる。これは、式(12)で表現されるエネルギー関数を減少させるように各ニューロンは動作する。

$$\frac{dv_i}{dt} = \sum_j T_{ij}x_j + I_j \tag{10}$$

$$x_i = \frac{1}{1+exp(-\frac{v_i}{\lambda})} \tag{11}$$

$$E = -\frac{1}{2}\sum_i\sum_j T_{ij}x_j - \sum_i I_ix_i \tag{12}$$

そこで、ニューロン間の相互の重み$T_{ij}(T_{ij}=T_{ji}, T_{ii}=0)$および、外部入力$I_i$を決めることにより、任意のエネルギー関数に対し、極小点を求めることができる。しかし、エネルギー関数は、一般に複数の異なる極小点をもち、常に最小点を得られるというわけではない。

式(9)をホップフィールドニューラルネット上に写像し、解を求める。すなわち、式(9)と式(12)の係数を合わせれば、ホップフィールドニューラルネット上で解くことができる。

5 むすび

ニューラルネットワークによるデータマイニングについて考察した。特徴は次の通りである。

1) データベース内に互いに競合する無矛盾集合が複数存在する場合、それらの中から極大なものを求めるのに適している。
2) ニューラルネットワークによる近似解法は並列処理向きである。
3) ルール表現を直接的に求める手法を提供する。ただし、ルールはシングルクローズ形式である。
4) ルール表現は"アナログ的な表現"を許しており、人間の知識との整合性がとりやすい。

現在4)に着目し、人間の知識との整合/検証を行なうシステムについて考察を行なっている。

参考文献

[1] Rakesh Agrawal, Arun Swami,"Database Mining: A Performance Perspective", IEEE Trans. on Knowledge and Data Engineering, Vol.5, No.6, 1993.

[2] Takahiko Shintani, Masaru Kitsuregawa," Consideration on Parallelization of Database Mining", 信学技報 DE95-74,pp.57-62,1995.

[3] 久間 和生, 中山 高,"ニューロコンピュータ工学", 工業調査会,1992.

[4] J.J.Hopfield and D.W.Tank ,"Neural Computation of Decisions in Optimization Ploblems" ,Biological Cybeenetics,52,pp.141-152,1985.

[5] 福本 光, 樋渡 咲, 呉 旭, 阿江 忠,"ニューラルリーズニングによる知識獲得", 人工知能学会研究会資料,SIG-PPAI-9503,pp.32-37,1996.