アンサンブル学習を用いたデータマイニングによる意思決定支援

日大生産工(院)　○町田和明
日大生産工　　山下安雄

1. はじめに

近年、大容量記憶媒体の低価格化、計算機処理能力の向上、情報通信技術の急速な発展に伴い、様々な分野で多量のデータを扱うことが多くなってきている。そのような中、大規模データからのデータマイニングによる有用な知識獲得の必要性がより求められるようになってきている。そして、これに付随する形で、金融分野、流通･小売分野、製造分野、通信分野、製薬･医療分野といった様々な分野において、現在使えるデータマイニング技術を用いた実社会への適用事例も報告されるようになってきている。このように、データマイニングに対する期待は大きく、今後、益々発展していくことが予想される。

そこで、本研究では、UCI が提供する国勢調査のデータベースをもとに、アンサンブル学習という方法を用いて知識獲得((年間所得>$50k/yr)or(年間所得<=50k/yr))を行う。

2. アンサンブル学習

アンサンブル学習とは、一般に、与えられたデータから複数の仮説(分類器)を生成し、それらを適切に組み合わせて予測することにより、単一の仮説で学習した場合よりも予測精度を向上させようというものである。

この考え方を利用したものはいくつかあるが、その中でも今回は Bagging の手法を利用する。

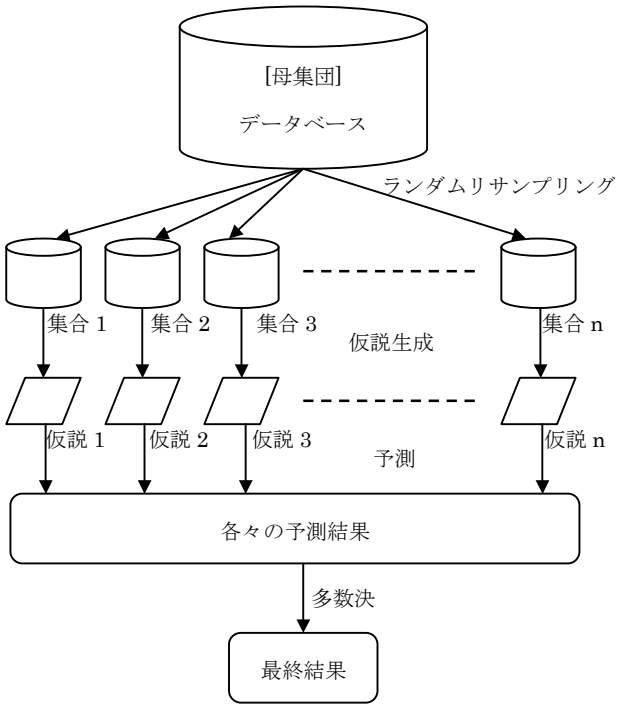

Fig.1　Bagging によるアンサンブル学習

Bagging は、データ集合を一様にランダムなリサンプリングをして、同サイズのデータ集合を複数個生成し、そのおのおのに同じ学習アルゴリズムを適用する。その後、それら複数回の試行により得られる出力仮説に対し、学習時データセットとは異なるテスト用データセットを用いて、その結果を多数決により最終的な予測とする(Fig.1)。

3．方法

3.1　データマイニングのプロセス

本研究では、UCI のデータベースより、あ

Support of Decision Making by Data Mining based on Ensemble Learning

Kazuaki MACHIDA and Yasuo YAMASHITA

Table 1　使用データ概要

Number of Instances
32561 (train=21708, test=10854)
Number of Attributes
14 (6 continuous, 8 nominal attributes)

Attribute Information

attribute	contents of attribute (extraction)
age	continuous
workclass(8)	Private, Self-emp-not-inc, …
fnlwgt	continuous
education(16)	Bachelors, Some-college, …
education-num	continuous
marital-status(7)	Married-civ-spouse, Divorced, …
occupation(14)	Tech-support, Craft-repair, …
relationship(6)	Wife, Own-child, …
race(5)	White, Asian-Pac-Islander, …
sex(2)	Female, Male
capital-gain	continuous
capital-loss	continuous
hours-per-week	continuous
native-country(41)	United-States, Cambodia, …

(注)属性名の後の( )内の数字は各々の属性において項目がいくつあるかを示す。また、( )の記述が無いものいついては、その属性の項目が連続量(整数値)であることを示す。

Class Distribution
2 (>50K, <=50K)

る年の国勢調査の結果をもとに抜粋されたデータセットを用いる。

アンサンブル学習にはこのデータセットを使用し、Bagging によりここから複数個の仮説を生成する。尚、これらの仮説は、アンサンブル学習を行う上での下位層に位置することから、下位学習アルゴリズムを呼ぶものとする。そして、今回この下位学習アルゴリズムには、BP ネットワーク(誤差逆伝播アルゴリズム)を用い、最終結果を多数決により導出する。

3.2　データセット仕様

使用するデータは、UCI が提供する数多くのデータベースの中から、国勢調査に基づき年収を 2 クラス分類するという趣旨でまとめられたデータセットを用いる。このデータセットは、1994 年の国勢調査のデータベースをもとに、Barry Becker 氏がある条件の下に抜粋したものである。

インスタンスの数は 32561 で、そのうちの 2/3(21708)は訓練用、1/3(10854)はテスト用となっている。また、各インスタンスにおける属性は 14 あり、整数型と分類型のいずれかによって 1 レコード(インスタンス 1 つ)が生成されている。尚、各レコードには、そのインスタンスの年収がどのクラス(2 クラスのうちどちらか)に属するかという情報も付加されており、これは学習時の教師出力となる情報であるといえる(BP ネットワークは教師あり学習のためこの情報は必要不可欠)。上記をまとめたものを Table 1 に示す。

3.3　学習のためのデータ前処理

3.1 で示したように、本研究では下位学習アルゴリズムには BP ネットワークを使用する。また、データセットには、3.2 で示したように、各レコードの項目(属性)が１４あり、かつ、それらの属性は整数型と分類型があるものを使用する。ここで問題となるのが、これらの情報をどのように学習させるかということである。このままの状態では、属性の型が統一されていないこと、また、整数型における数値の範囲が属性によって大きくばらつくということが問題であるといえる。

そこで、これらばらつきのあるデータ属性を一定の範囲内に正規化する処理が必要になると考えられる(データにばらつき(偏り)があると、値の大きなもの(ニューロン)に引っ張られて、小さな値の影響がかき消されてしまう可能性がある)。したがって、今回は、分類型の属性に関しては、各々の属性数に応じたぶんだけニューロンの数を用意し、それぞれのカテゴリに一個のニューロンが発火するという形式をとる。また、整数型の属性に関しては、カテゴライズされていないため、属性ごとの最大値および最小値よりその範囲を数等分してカテゴリに分け、後は分類型と同様、ニューロンに

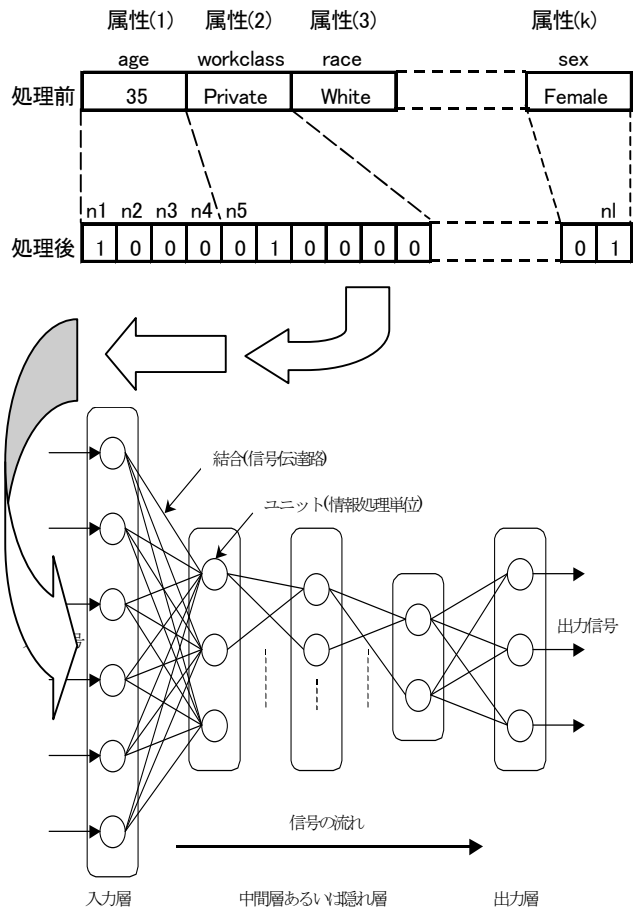

Fig.2　データの前処理から BP ネットワークへの適合プロセス

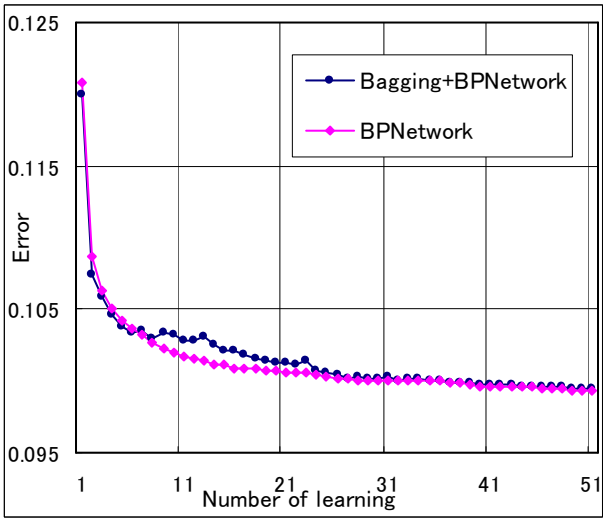

Fig.3　各学習器の学習過程

Table 2　ネットワーク出力分布

		Output of network			
		>50		<=50	
Output of teacher	>50	(B+B)	1690	(B+B)	741
		(B)	1462	(B)	631
	<=50	(B+B)	966	(B+B)	7457
		(B)	1194	(B)	7567

Table 3　認識率・誤認識率

Recognition rate (%)	(B+B)	84.27
	(B)	83.18
Misrecognition rate(%)	(B+B)	15.73
	(B)	16.82

一個ずつ対応させていき、最終的に 0～1 の範囲(0 or1)に正規化された BP ネットワークへの入力データセットを得るものとする。

この前処理の様子をまとめたものを Fig.2 に示す。

4. 結果

訓練用のデータセット、及びテスト用データセットを用いて学習・識別を行った結果を Fig.3、Table 2、Table 3 に示した。

Fig.3 は、３．方法で示したアンサンブル学習とその下位学習アルゴリズムの BP ネットワーク単体での、訓練用データセットに対する学習過程を示してものである。これは、横軸が学習回数(単位;エポック×10)、縦軸がエラーである。このエラーというのは、０～１の範囲を持ち、0 に近いほど教師出力と実際のネットワークの出力との差が無いことを示す。

また、Table 2 及び Table 3 は、Fig.3 で学習させたネットワークにおいて、テストデータを用いて識別を行ったものである。Table 2 では、ネットワークが実際の教師出力のカテゴリに対してどのような出力を呈したかということを示してある。尚、図中の B+B 及び B は、それぞれ Bagging＋BP ネットワーク、BP ネットワークであることを示している。そして、Table 3 は、Table 2 において、それぞれ正しく分類されたもの、間違って分類されたものの確率を示した。

以上により、BP ネットワーク単体のものよ

り、アンサンブル学習を行ったほうが少し認識精度が向上することがわかった。

5. まとめと今後の展望

今回は、Bagging の手法を用いたアンサンブル学習において、下位学習アルゴリズムに BP ネットワークを取り入れた。結果に示したように、アンサンブル学習の方が少し良い結果にはなったが、今回用いたアンサンブル学習の手法は最も簡単(単純)なものであり、かつ、下位学習アルゴリズムに BP ネットワークというひとつのパターンでしか検証していないので、更なる認識精度向上のためには一工夫必要であると考えられる。

ここで考えられるものの一つに、下位学習アルゴリズムに他の学習アルゴリズムを導入するということである。

例えばこれには、決定木学習、決定木学習とは異なった原理で動作するルール学習、確率の考え方に基づくナイーブベイズ学習、過去のデータと未知のデータとの距離を計算して未知データのクラスを推定する最近傍法、頻度の考え方に基づき大量データに対応できる相関ルールマイニングなど、実に様々な手法が考えられる。

また、もうひとつの考え方として、下位学習アルゴリズムではなく、アンサンブル学習の方法そのものを変えるということである。今回は、Bagging をアンサンブル学習の手法として用いたが、この外にも、ブースティングや確率的属性選択といった手法がある。また、更に、これらいろいろなアンサンブル学習をうまく組み合わせることで、その学習アルゴリズムの手法は実に多岐にわたる。

このように、考えられる学習アルゴリズムは多くあるので、多くの学習器を作成して比較検討してみたい。尚、これらの学習アルゴリズムをただ使えばよいというものではないということに注意が必要である。なぜなら、これら学習アルゴリズムのいずれもが優れた結果を示すとは限らないからである。それぞれの学習アルゴリズムには特徴があり、学習させようとするデータの特徴(構造)によってもその向き不向きがあるのが現実である。したがって、まず、対象とするデータの特徴をよく吟味して、その上で、ある程度効果的と思われるものを選択していくことが大切であると思われる。

以上より、今後は、使用するデータ特性に合いそうないくつかの学習アルゴリズム(下位学習アルゴリズムならびに(上位の)アンサンブル学習アルゴリズム)を用いて比較検討することで、機械学習ならびにデータマイニングによる知識獲得へと展開していきたいと考える。

「参考文献」

１） 馬見塚拓，安倍直樹，集団能動学習 ―データマイニング･バイオインフォマティクスへの展開―，電子情報通信学会誌，Vol.J85-D-Ⅱ，No.5，(2002)，pp.717-724.

２） 藤原由希子，富沢伸行，井口浩人，閾値関数による変化分析型アンサンブル学習を用いた商品･サービス普及予測，第 7 回情報科学技術フォーラム講演論文集，(2008)，pp.23-25(第 2 分冊).

３） R.Kohavi and B.Becker，UCI Machine Learning Repository Adult Data Set(1996), http://archive.ics.uci.edu/ml/datasets/Adult

４） Simon Haykin，Neural Networks，Macmillan College Publishing，(1994)

５） 元田浩，津本周作，山口高平，沼尾正行，データマイニングの基礎，オーム社，(2006)

６） C.M.ビショップ(元田浩 他監修)，パターン認識と機械学習，シュプリンガー･ジャパン，(2007)