2002年日本オペレーションズ・リサーチ学会
春季研究発表会

データベースからの重要属性の抽出

甲南大学　*小寺　崇弘　KOTERA Takahiro
甲南大学　中山　弘隆　NAKAYAMA Hirotaka

1. はじめに

データベースからの重要属性の抽出はエキスパートシステムにおいて重要な事柄の1つである。重要属性の抽出において、よく用いられる指標として整合度と情報量(エントロピー)がある。整合度は矛盾していないデータの割合を表し、情報量ではデータベースが与える情報量と集合を分類した場合の情報量の差を表す。この2つの指標では、重要属性の抽出に若干の違いが見られる。本研究では、2つの指標の違いを検討する。また重要属性の選択方法についての考察も行う。

2. 決定表

表1のようなデータの集まりを決定表と呼び、条件属性のとりうる値によって決定属性値が決まるものである。属性1の値が0である決定属性値の割合は1:2、属性1の値が1である決定属性値の割合は3:1になっている。また属性2の値が0である決定属性値の割合は2:2、属性2の値が1である決定属性値の割合は2:1となっている。割合から見ると、属性2の方が矛盾の度合いが高いことが分かる。この表1より属性1と属性2のどちらが重要な属性であるか2つの指標を用いて検討をする。

表 1: 重要属性を抽出する決定表

	属性1	属性2	決定属性
1	0	0	1
2	0	0	1
3	0	0	0
4	1	0	0
5	1	1	0
6	1	1	0
7	1	1	1

3. 整合度

重要属性の抽出を行うために、データベースの整合度を考える。これは、どれだけ矛盾したデータが含まれているかの割合によって表される。この値は抽出した属性がどれだけ矛盾のない質の高い知識を含んでいるかの1つの指標となる。データベースの整合度は、

$$\gamma_R(\chi) = \frac{\sum_{i=1}^{n} card(\underline{R}X_i)}{card(U)} \tag{1}$$

によって表される。データベースの整合度は、0から1までの値をとり、1に近づく程矛盾のないデータベースであると言える。表1における属性1の整合度は、$\gamma_R(\chi) = 0.00$ となる。また属性2の整合度も、$\gamma_R(\chi) = 0.00$ となる。この結果より、属性1、2ともに整合度は0.00になり、2つの属性の重要度に差はないという判断をすることができる。

4. 情報量(エントロピー)

重要属性の抽出を行うためには、決定表が与える情報量と抽出した条件属性が与える情報量の差を求める。情報量の差が大きいほど、抽出した条件属性が重要であることが分かる。

step 1:集合 C において決定表が与える全情報量を求める。

$$M(C) = -\sum_{i=1}^{L} \frac{n_i}{n} log_2 \frac{n_i}{n} \quad [bit] \tag{2}$$

step 2:抽出した属性 A によって集合を分類した場合に、決定表が与える情報量を求める。

$$B(C,A)=\sum_{i=1}^{L}p(C_i)\times M(C_i) \tag{3}$$

step 3:抽出した属性 A に対し、分類によって得られる情報量を求める。

$$M(C)-B(C,A) \tag{4}$$

表 1 における属性 1 の情報量は、$0.128[bit]$ となる。また属性 2 の情報量は、$0.020[bit]$ となる。この結果より、属性 1 の方が重要な属性であることが分かる。整合度だけでは、判別できなかった微妙な重要性を情報量は判別できたということになる。

5. 重要属性の抽出

重要属性の抽出には、3 種類の方法がある。

1) 全ての条件属性の組合せを考える。

2) 不要な条件属性を減らす。

3) 必要な条件属性を増やす。

確実に重要属性が抽出できるのは、全ての組合せを考えるときである。しかし条件属性数が多くなるにつれ、組合せの数も増えていくので処理時間が長くなってしまう。それに比べ、不要な条件属性を減らす方法や、必要な条件属性を増やす方法では、組合せの数が少ないため短時間で処理できる。例えば倒産予測問題 (R.Slowinski *et al*. 1995; 条件属性 12 個) の場合に必要属性を 4 個に絞ると、全ての組合せは 495 通りあるが、上記 2,3) の方法では、42 通りしかない。とくに、3) の方法では、全ての条件属性の組合せを考えた場合と全く同じ結果が得られている。

表 2:全ての条件属性の組合せ

使用条件属性				総合情報量
1	3	7	9	1.485731
1	3	7	11	1.485731
1	3	7	12	1.485731
1	7	8	9	1.485731
1	7	8	11	1.485731
2	3	6	7	1.485731
3	6	7	8	1.485731
3	6	7	9	1.485731
3	6	7	12	1.485731
6	7	8	11	1.485731

表 3:不要な条件属性を減らす

使用条件属性				総合情報量
7	6	9	2	0.846154

表 4:必要な条件属性を増やす

使用条件属性				総合情報量
7	8	11	1	1.485731

6. 終わりに

組合せの数を減らしても、以前と変わりなく重要属性を抽出することができた。また斜面崩壊予測問題 (データ数:53369、条件属性:22) に関しても同じように重要属性が抽出できている。今後は、どのデータベースにも適応できるものか確認する必要がある。

参考文献

[1] Pawlak,Z : Rough Sets.Theoretical Aspects of Reasoning about Data.Kluwer Academic Publishers,Dordrecht.(1991)

[2] J.R. キンラン : AI によるデータ解析, トッパン (1995)

[3] R.Slowinski and C.Zopounidis:Application of the Rough Set Approach to Evaluation of Bankruptcy Risk, Intelligent Systems in Accounting, Finance and Management vol.4:27-41 (1995)