CS-82

条件付きエントロピーによるデータマイニングの基礎検討

武蔵工業大学　学生会員　高須 光郎
（株）地崎工業　正 会 員　須藤 敦史
武蔵工業大学　正 会 員　星谷 勝

1．はじめに

マーケティング分野において，大規模なデータの高度な処理を行うデータマイニング [1]の数々の適用事例が発表されている．データマイニングは，膨大なデータの中に存在する隠れた法則や相関関係を自動的に発見する手法である．そして，定性および定量データが混在するデータや属性の相関関係が複雑なデータも扱うことが可能であり，その適用分野を問わないのが特徴である．加えて，土木分野においてもデータマイニングは情報分析手法として対応でき，その有用性も期待される．そこで本研究はデータマイニングの適用を目的とし，その概要や適用例を示すとともに，データマイニングにおいて重要な要素となる条件付きエントロピーの検討を簡単な数値解析を通して行っている．

2．データマイニング

データマイニングの名称は鉱山から鉱脈を採掘することに由来し，93 年に米 IBM アルマデン研究所のアグラワラ研究員が提唱したものである．従来の解析手法はデータが膨大になると分析が難しくなり，細部に至るまでは人間の目が行き届かず，データが十分に活用できないことがある．しかし，データマイニングは人間では想像もできないような規則性を見つけ出し，データ中に存在する情報を引き出すことが可能である．さらに分析作業を自動化することにより，分析結果の客観性と信頼性を高め，エンドユーザーに高度な意思決定を支援することができる．

適用事例（マーケティング） あるスーパーマーケットにおいて売上実績データをデータマイニングツールで処理し，顧客の購入心理とその規則性について分析した結果，予想以上に多くの既婚男性客がビールと紙おむつを同時に買っていくパターンを発見した．そこでビールと紙おむつの売場を並べてディスプレイしたところ，売上げの増加につながった例が挙げられる．

データマイニングにおける事象の関係「IF～，THEN …」形式が「事象 a_i が与えられたときの事象 b_j の条件付き確率」に対応している．ここで本研究は情報エントロピーを導入し，条件付き確率と情報エントロピーがデータマイニングの手段としてなりうることを検討する．

3．条件付きエントロピー

エントロピー [2]とは「不確定の度合い」といった極めて抽象的なものを定量的に評価するものであり，ある事象のエントロピーはその確率分布によって一意的に定まる．離散型確率分布 $P(A=a_i)=p_i\ (i=1,\cdots,m)$ を有する事象 A のエントロピーは式(1)のように与えられる．

$$H(A)\equiv H(p_1,\cdots,p_m)=-\sum_{i=1}^{m} p_i \log p_i \qquad (1)$$

次に，確率分布 $P(B=b_j)=q_j\ (j=1,\cdots,n)$ を有する事象 B が起こったときの，A の条件付きエントロピーは式(2)のようになる．

$$H(A|b_j)=H(P(a_1|b_j),P(a_2|b_j),\cdots,P(a_n|b_j)) \qquad (2)$$

ここで，$P(a_i|b_j)$ は b_j が発生したときの a_i の条件付き確率であり，$H(A|b_j)$ は b_j が発生したときの A の条件付きエントロピーである．これは A と bj が独立のときは $H(A)$ と一致し，，相関がある場合には $H(A)$ に比べ減少する．

KEYWORD　データ処理　確率論　条件付き確率　情報エントロピー
連絡先　住所　〒158-0087 世田谷区玉堤 1-28-1　TEL 03-3703-3111(ex.3268)　FAX 03-5707-2187

4．数値計算例

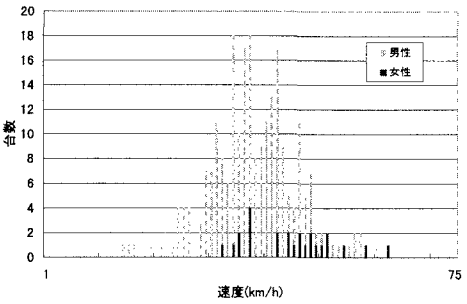
図-1から図-3に示す3つの（交通）データSampleⅠ,Ⅱ,Ⅲを用いて数値解析を行った．この図は運転手の性別Aと速度Bについて，男性215台と女性25台（計240台）に分けてヒストグラムで表したものである．

これら3つのデータは，男女合せた全体の総数（240台）とその分布形は同一である．異なる点は，男女それぞれの分布形である（ただし，データ数は同じである）．SampleⅠは女性分布がばらついているもの，SampleⅡは女性分布が寄せ集まっているもの，SampleⅢはSampleⅡと同一の女性分布が右に寄っているものである．もちろん男性分布もそれぞれ異なっているが，その特徴についてはここでは省略する．表-1に男女合せた全体の速度のエントロピーH(B)，男性であることが確認されたときの速度のエントロピー$H(B|a_1)$，女性であることが確認されたときの速度のエントロピー$H(B|a_2)$，性別（男女どちらかはわからなくてよい）が確認されたときの速度のエントロピーH(B|A)を示す．H(B|A)は$H(B|a_1)$と$H(B|a_2)$の重み付き平均である．

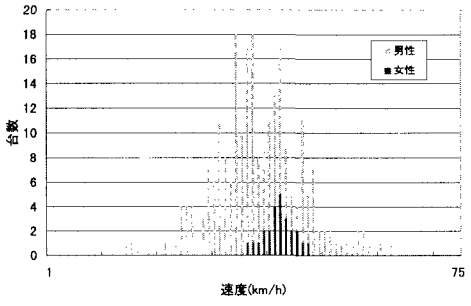
SampleⅠ～SampleⅢすべてにおいて，H(B|A)がH(B)に比べ減少している．これより，性別Aと速度Bは独立なものではなく，何らかの相関があることがわかる．またその相関の度合いは，H(B)とH(B|A)の差よりSampleⅡが最も弱く，以下SampleⅠ，SampleⅢの順に強くなる．

次に，H(B|A)の要素である$H(B|a_1)$と$H(B|a_2)$の大小関係について考察する．SampleⅠ～SampleⅢすべてにおいて，$H(B|a_1) > H(B|a_2)$となっている．これより，男性に比べて女性のほうが速度との相関が強いことがわかる．

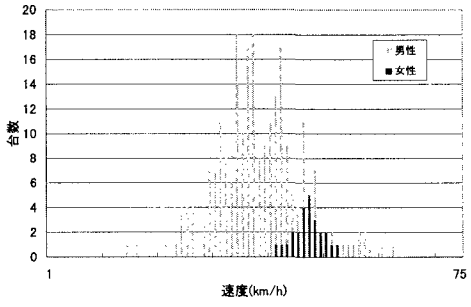
最後に，$H(B|a_2)$の変化について考察する．SampleⅠとSampleⅡについて比較するとSampleⅡのほうが小さい．これより，女性分布がSampleⅠのばらついたものよりSample-Ⅱのまとまったもののほうが$H(B|a_2)$が小さくなることがわかる．ところが，SampleⅡとSampleⅢについて比較すると$H(B|a_2)$に変化が見られない．これより，女性分布の平均値は$H(B|a_2)$には影響しないことがわかる．つまり，$H(B|a_2)$は女性の分布形のみに依存し，平均値には依存しないのである．



図－1 SampleⅠ



図－2 SampleⅡ



図－3 SampleⅢ

表－1 条件付エントロピーの比較

	H(B)	$H(B\|a_1)$	$H(B\|a_2)$	H(B|A)
SampleⅠ	3.36055	3.29113	2.66436	3.22584
SampleⅡ	3.36055	3.37211	2.32154	3.26267
SampleⅢ	3.36055	3.24782	2.32154	3.15133

a_1:男性，a_2:女性

5．まとめ

将来，データマイニングは土木分野の中でも多種多様な応用範囲が考えられる．今後は実データに適用する予定である．

＜参考文献＞

1). 情報・通信新語辞典97年度版，日経ＢＰ社，1996　　2). 有本卓：確率・情報・エントロピー，森北出版，1992