5M-6　地上気象観測データからの相関ルール抽出

白石 將,　田中 秀俊,　木瀬 若桜
三菱電機 (株) 情報技術総合研究所

1. はじめに

気象予測に利用可能な知識の獲得を目的として，気象データからの相関ルール抽出実験を行った．本解析で対象とする相関ルールは「$A_1,\cdots,A_n \leftrightarrow B$」の形式をしており，「アイテム集合 $A_1,\cdots,A_n$ とアイテム B は同時に生起することが多い」ことを表す．便宜上，相関ルールの左辺 $(A_1,\cdots,A_n)$ を条件部，右辺 (B) を結論部と呼ぶ．

解析対象としたデータは気象庁提供の地上気象観測データであり，夏期の夜間における釧路の霧発生や視程悪化などの事象に対して強い相関を持つ気象事象を抽出することを目的とした．最終的にやりたいのは気象予測なので，「ある事象が生起してしばらくして別の事象が生じる」ことを示唆するような，時間差を含む相関ルールを抽出できるよう，事前にデータを加工した上で相関ルール抽出アルゴリズムを適用した．本稿では，解析方法およびその結果について述べる．

2. 解析手順

解析対象のデータは気象庁提供の 1991 ～ 1998 年の 6 ～ 8 月分の地上気象観測データ [1] である．本解析では釧路の気象事象に対して強い相関を持つ気象事象を抽出することが目的であるので，釧路および，釧路と近い広尾および根室のデータを選択して用いた．

次にデータの内容について説明する．解析に利用したのは時別値が格納されているものであり，毎日の 1,2, …, 24 時における現地気圧, 気温, 露点温度, 蒸気圧, 風向, 風速, 降水量などのデータが記録されている．また，釧路については毎日の 3,6,9,12,15,18,21 時における天気や視程などのデータが記録されている．

以上のデータに対して適用した解析手順を以下に示す．

1. 解析対象のデータに存在する属性を組み合わせて，相関ルールに現れた際に理解が容易であるような新たな属性を作成する．例えば空気の湿り気を表現する指標として「気温と露点温度の差」，また霧状態の開始・終了を表す指標として「気象変化」属性を作成した．
2. 連続する 6 時点分のデータを 1 つのまとまり (以下レコードと呼ぶ) に変換する処理を行う．連続する 6 時点は，その第 6 時点が 18,21,3,6 時のいずれかとなるように選択した．これは，これらの時刻において釧路の天気・視程データが取得されており，また本解析の目的が釧路での夜間の霧発生や視程悪化に関する知識を得ることであるためである．6 時点分のデータの選択後，以下の属性値を取り出して並べることにより，1 つのレコードを生成する．
   - 第 4 時点のデータ．
   - 第 3 時点から第 4 時点にかけてのデータの 1 時間分の時間変化．
   - 第 1 時点から第 4 時点にかけてのデータの 3 時間分の時間変化．
   - 第 6 時点の天気, 視程, 気象変化の値 (釧路のデータのみに適用).

   ここで，1,2,3 番目の項目は相関ルールの条件部に，また 4 番目の項目は結論部に出現させることを想定している．このようにすることにより，釧路の天気や視程に関する気象事象と，その 2 時間前までに得られているデータとの相関を表すルールを得ることが可能となる．
3. 釧路, 広尾, 根室のデータにおいて，対応する時点のレコード同士を 1 つのレコードにまとめる．
4. 連続値属性の取り得る範囲を適切な数の境界値で区切り，各領域に新たな名前をつけて新属性値とする．
5. 不必要と判断される属性や属性値に関するデータを削除する．
6. 属性値に属性名を連結することにより，相関ルールの構成要素となるアイテムに変換する．以上までの処理を実行することにより，321 種類のアイテムからなる 2,944 個のレコードが生成された．
7. アイテムのレコードへの同時出現の頻度を数え上げていくことにより，出現頻度が高く，また出現の相関が強いようなアイテムの組を探索し，相関ルールとする．ここで相関の強さの指標としては χ^2 値を用いる [2]．以下の条件を満たす相関ルールを探索したところ，121,255 個の相関ルールが得られた．
   - 頻度が 100 以上，χ^2 値が 30 以上．
   - 釧路の天気, 視程, 気象変化に関する第 6 時点のアイテムが結論部のみに出現，それ以外のアイテムが条件部のみに出現する．
8. 最後に，χ^2 値を基準とした不要相関ルール削除を行う．要不要の判断基準は以下の通り．ある相関ルールに対して，条件部にさらにアイテムを付加して生成された相関ルールを考える．アイテムを付加して制限を強くしたのにもかかわらず，相関の強さを示す指標である χ^2 値が増加しなかった場合，生成された相関ルールの χ^2 値はもとの相関ルールの χ^2 値を反映しているに過ぎないと判断して削除する [3]．この不要相関ルール削除を実施した結果，783 個の相関ルールが残った．

Extracting Association Rules from Surface Meteorological Databases
Masashi SHIRAISHI, Hidetoshi TANAKA, Wakasa KISE
Mitsubishi Electric Corporation

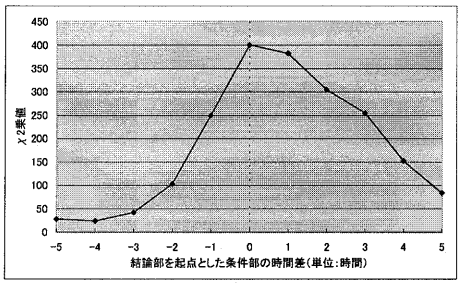

図 1: 時間差と χ^2 値との関係 (相関ルール A)

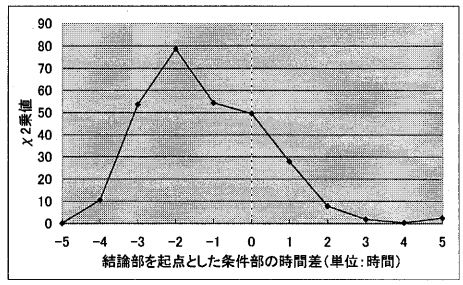

図 2: 時間差と χ^2 値との関係 (相関ルール B)

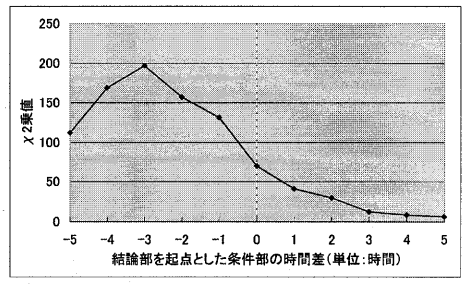

図 3: 時間差と χ^2 値との関係 (相関ルール C)

3. 得られた相関ルールに関する考察

実際に得られた相関ルールの例を以下に示す．但し，確信度とは「条件部が生起した場合にさらに結論部が生起する割合」を表す数値である．

A 釧路の降水量が 0 であり，かつ気温と露点温度の差が 1 以下であるという事象 (条件部) は，その 2 時間後に釧路で霧状態が開始するという事象 (結論部) と強い相関がある (χ^2 値は 103.1，確信度は 21%).

B 広尾の降水量が 0 であり，かつ根室での 3 時間分の現地気圧の変化が -2.3 から -0.3 の範囲にあるという事象 (条件部) は，その 2 時間後に釧路で霧状態が開始するという事象 (結論部) と強い相関がある (χ^2 値は 78.7，確信度は 17.4%).

C 釧路の気温と露点温度の差が 1 以下であるという事象 (条件部) は，その 2 時間後に釧路で霧状態が終了するという事象 (結論部) と強い相関がある (χ^2 値は 157.4，確信度は 12.6%).

このように，釧路の霧や視程悪化等の気象事象 (結論部に相当) と，その 2 時間前までに得られているデータ (条件部に相当) との相関を表すルールが得られている．しかし，「2 時間後」という条件部と結論部の時間差を表す数値に本当に意味があるのかどうかについては明らかではない．そこで，各相関ルールについて，その時間差を変えることによって得られる相関ルールの χ^2 値を算出するプログラムを作成し，検証を行った．

上に示した相関ルール A,B,C に関し，結論部を起点とした条件部の時間差を -5 から 5 まで変化させた場合の，対応する相関ルールの χ^2 値を示すグラフを図 1,2,3にそれぞれ示す．ここで，負の時間差は条件部が結論部より時間的に前である場合を，また正の時間差は条件部が結論部より時間的に後である場合を表す．

これらのグラフより，以下の事柄が推論できる．

- 図 1によれば時間差が 0 のところで χ^2 値が最大となっている．従って相関ルール A は，実は「条件部と結論部が同時に生起することが統計的に見て多い」ことを反映したものであり，結論部の予測に利用できるようなルールではない可能性が大きい．
- 図 2によれば，時間差が -2 のところで χ^2 値が最大となっている．つまり相関ルール B は「条件部が生起して 2 時間後に結論部が生起することが統計的に見て多い」ことを表しており，結論部の予測に利用できるようなルールである可能性がある．
- 図 3によれば，時間差が -3 のところで χ^2 値が最大となっている．つまり相関ルール C は「条件部が生起して 3 時間後に結論部が生起することが統計的に見て多い」ことを表しており，結論部の予測に利用できるようなルールである可能性がある．

しかし，相関ルール C の場合，「条件部が霧状態終了の前兆となっている」という解釈は実は正しくなく，単に「条件部は (霧状態終了前における) 霧状態と同時に生起することが多い」ことを表しているに過ぎないと考えるのが自然である．この例からもわかるように，時間差を含む相関ルールにおいては，その解釈を注意深く行う必要がある．

相関ルール B についても，「統計的に相関が強いアイテムの組」を示しているのであって，因果関係が実際に存在するかどうかの確認は別途行う必要がある．また，この相関ルールは確信度が 17.4% と低いために直接予測に利用することはできないが，気象変化に関するモデルの構築や，予測する際に利用する変数の組の選択などに用いることができると考えられる．

4. おわりに

地上気象観測データからの，時間差を含む相関ルール抽出の方式および結果を示した．今後，対象とするデータの種類を増やすなどして実験を進める予定である．

参考文献

[1] 気象庁監修: 地上気象観測原簿データ CD-ROM6 枚 (1989 年 4 月～ 1998 年分).

[2] 三石他: Knodias におけるデータマイニング方式, 第 56 回情報処理学会全国大会 2W-6 (1998).

[3] 川上他: 相関係数による数理的フィルタリング機能検証, 第 10 回データ工学ワークショップ [DEWS'99] (1999).