相関ルール (Association Rule)

相関ルール (Association Rule)

相関ルールとは, X ならば Y であるという形式で表されるルール. このルールを発見する分析をアソシエーション分析と呼ぶ. アソシエーション分析に伴う困難は, 発生する事象全てをルールとして抽出したのでは, 大きなデータに対しては必然的に膨大な量のルールを見つけてしまう事にある. そこでルールの有益度を測る指標としては支持度, 信頼度, リフト値という指標が良く使われる.

評価指標

代表的な評価指標として支持度 (Support), 確信度 (Confidence), リフト値 (Lift) が挙げられる. データ項目の集合 $I = \{A, B, C, \dots\}$, トランザクションは I の部分集合, $T(X)$ は X を含むトランザクション集合とする.

支持度 (Support)

ある相関ルール $(X \Rightarrow Y)$ がどれだけ役立つかを表す指標. 全データにおける $X \cup Y$ が占める割合つまり, $X \cup Y$ が起きる確率. $X \cup Y$ を含むデータが多いほど相関ルールは支持される. D は全データの件数とする.

$$S = \frac{|T(X) \cap T(Y)|}{D} = \frac{|T(X \cup Y)|}{D}$$

確信度 (Confidence)

ある相関ルール $(X \Rightarrow Y)$ において X と Y の相関の強さを表す指標. X に続けて Y が起こる割合 (条件付確率).

$$C = \frac{|T(X) \cap T(Y)|}{|T(X)|} = \frac{|T(X \cup Y)|}{|T(X)|}$$

リフト値 (Lift)

ある相関ルール $(X \Rightarrow Y)$ において Y の確率が高い場合の確信度は Y によるものつまり, 相関ルールとして Y の確率に X がどのように影響したかを計る指標が必要となる. それがリフト値である. ある相関ルール $(X \Rightarrow Y)$ のリフト値は X による Y の確率の上昇率を表す. Y の確率 $P(Y)$ とする. X が起きたときの Y の確率 (条件付確率) を以下のようにする.

$$P(Y|X) = \frac{|T(X \cup Y)|}{|T(X)|} = \frac{P(X \cap Y)}{P(X)}$$

リフト値が 1 よりも大きいならば X と Y には正の相関がある.

$$L = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)} = \frac{C}{P(Y)}$$

アプリオリアルゴリズム (Apriori Algorithm)

支持度と確信度が高いほど相関ルールは有益である．すべてのルールを探索すると計算量が膨大になってしまう．そこで有益でないルールを除外するしきい値として最小支持度, 最小確信度を設定しそれを超えるような相関ルールを探査するアルゴリズムを考える．アプリオリアルゴリズムでは, $Support(P)$ が最小支持度よりも小さければ $P \subseteq Q$ となる $Support(Q)$ も最小支持度より小さいという性質を利用する．要素数の少ない集合から $Support$ を逐次的に計算し, 最小支持度を下回る集合と, それを含む集合を除外することで計算量を削減するアルゴリズムである．

データ項目の集合の $Support$ を求めれば, 支持度と確信度求められる．$Support(X) = \frac{|T(X)|}{D}$ とする．
$(X \Rightarrow Y)$ の支持度は

$$S = \frac{|T(X \cup Y)|}{D} = Support(X \cup Y)$$

$(X \Rightarrow Y)$ の確信度は

$$\begin{aligned} C &= \frac{|T(X \cup Y)|}{|T(X)|} \\ &= \frac{\frac{|T(X \cup Y)|}{D}}{\frac{|T(X)|}{D}} \\ &= \frac{Support(X \cup Y)}{Support(X)} \end{aligned}$$

$P \subseteq Q$ ならば $Support(P) \geq Support(Q)$ となる．

$$\begin{aligned} &Support(P) \geq Support(Q) \\ &\frac{|T(P)|}{D} \geq \frac{|T(Q)|}{D} \\ &|T(P)| \geq |T(Q)| \end{aligned}$$

ほかにも $Support$ は以下の性質をもつ

$$Support(P_1 \cup P_2) \geq k \Rightarrow Support(P_1) \geq k, Support(P_2) \geq k$$

$$Support(P) < k \Rightarrow Support(Q) < k \quad (P \subseteq Q)$$

chatGPT による試験対策問題

問題 1

データマイニングにおけるアソシエーションルール分析では、以下の評価指標がよく用いられます。それぞれの指標について、その定義を説明してください。

1. サポート (Support)
2. 信頼度 (Confidence)
3. リフト値 (Lift)

(回答) データ項目の集合を $I = \{A, B, C, \dots\}$, トランザクションは I の部分集合とする．ある事象 X のサポート

とは事象 X を含むトランザクション集合の要素数を全体のデータ数で割ったもの, 全体に対して X がどれだけ発生するかを表す. $T(X)$ は X を含むトランザクション集合, D はトランザクション集合の要素数とする.

$$S(X) = \frac{|T(X)|}{D}$$

ある相関ルール $X \Rightarrow Y$ の信頼度とは X が発生したときにどれだけ Y が発生するかの条件付確率.

$$C(X \Rightarrow Y) = \frac{|T(X \cup Y)|}{|T(X)|}$$

ある相関ルール $X \Rightarrow Y$ のリフト値とは X が発生したときにどれだけ Y が発生する確率が上昇するかを表す指標. 信頼度だけではもともと Y の発生確率が高い場合に適切な判断ができないので, X の Y 影響度を測るために利用される. リフト値が 1 より大きい場合は X, Y には正の相関がある.

$$\begin{aligned} L(X \Rightarrow Y) &= \frac{\frac{T(X \cup Y)}{D}}{\frac{|T(X)|}{D}\frac{|T(Y)|}{D}} \\ &= \frac{C(X \Rightarrow Y)}{\frac{|T(Y)|}{D}} \\ &= \frac{C(X \Rightarrow Y)}{S(Y)} \end{aligned}$$

問題 2

以下の取引データが与えられています。このデータを基にして、次の問いに答えてください。

取引 ID	商品 A	商品 B	商品 C	商品 D
1	○	○		○
2	○		○	○
3	○	○		
4		○	○	○
5	○	○	○	

表 1　取引データ

1. 商品 A と商品 B の「サポート値」を計算してください。
2. 商品 A → 商品 B というルールの「信頼度」を計算してください。
3. 商品 A → 商品 B というルールの「リフト値」を計算してください。
4. 商品 A → 商品 B というルールが有用なルールといえるか、理由を述べてください。

(回答 1)

$$S(A) = \frac{4}{5}$$
$$S(B) = \frac{4}{5}$$
$$S(A \cup B) = \frac{3}{5}$$

(回答 2)

$$C(A \to B) = \frac{3}{4}$$

(回答 3)

$$L(A \to B) = \frac{3}{4} / S(B) = \frac{3}{4}\frac{5}{4} = \frac{15}{16}$$

(回答 4) サポート値と信頼度は 0.5 よりも高いがリフト値が 1 を下回っているため, 有益なルールではあるが新規性や意外性のあるルールとは言えない.

問題 3

以下のルールのうち、リフト値が高いものを選び、その理由を簡潔に述べなさい。

1. ルール A: 商品 X $\to$ 商品 Y (リフト値 $= 1.2$)
2. ルール B: 商品 M $\to$ 商品 N (リフト値 $= 2.5$)
3. ルール C: 商品 P $\to$ 商品 Q (リフト値 $= 0.8$)

(回答) ルール B. ある相関ルール $X \Rightarrow Y$ のリフト値は 1 を超えると X, Y に正の相関がある. よって最もリフト値の高いルール B が選ばれる.