5N-05

時系列パターンとアイテムセットの出現順序を考慮した分類パターンによる分類モデルの精度向上に関する一考察

小柳　暁奨　　新谷　隆彦　　大森　匡　　藤田　秀之
電気通信大学大学院　情報システム学研究科

1. はじめに

今日の技術進歩によって膨大なデータが取得されるようになり、データを有効活用するための研究が行われてきた。その一つにパターンを用いてクラス分類を行う分類パターンマイニング[1]がある。これまでに、多種類のデータを組み合わせた分類パターンによって分類精度を高める研究[2]が行われきたが、データ種類間の出現順序を考慮していなかった。データにタイムスタンプが含まれている場合、データ種類間の出現順序を考慮して分類パターンを抽出することが可能であり、より精度の高い分類パターンマイニングが期待できる。

そこで本研究では、時系列パターンとアイテムセットを出現順序を考慮して結合した分類パターンによる分類パターンマイニングにより、分類精度の向上を目指す。

2. 分類パターンマイニング

分類パターンマイニングとは一件のレコードがユーザ ID、アイテムからなるデータ、クラス値からなるデータベースから、アイテムで構成されるパターン(分類パターンと呼ぶ)を見つけ出し、その分類パターンを前提条件、クラス値を結論とするルールのリストを分類モデルとし、クラス値が未知のレコードにルールを上から順に適用してクラス値を予測する。

本研究ではタイムスタンプを持つデータであるシーケンスデータを処理対象とする。シーケンスデータはユーザ毎のアイテムがタイムスタンプ順に並べられたリストであるため、時系列パターンマイニングを適用することによって出現順序を考慮したアイテムのリストを抽出することが出来る。また、順序を無視することによってアイテムセット(出現順序を考慮しないアイテムの組合せ)を抽出することも出来る。

Consideration on mining class association rules using itemset and sequential pattern for improving classification accuracy
Akihiro Koyanagi, Takahiko Shintani ,
Tadashi Ohmori, Hideyuki Fujita
Graduate School of Information Systems
The University of Electro-Communications

表 1　データベースの例

ID	シーケンスデータ	クラス
1	(A,1),(B,2),(C,3),(D,4)	Y
2	(C,1),(D,2),(A,3),(B,4),(D,5),(C,6)	Y
3	(D,1),(C,2),(A,4),(B,5),(D,7),(C,8)	Y
4	(C,1),(D,3),(A,4),(B,5),(D,6),(C,7)	N
5	(B,1),(A,2),(D,3),(B,5),(C,7)	N
6	(B,1),(A,4),(C,5),(B,6),(D,7)	N

表 1 にデータベースの例を示す。このデータベースはユーザ数 6 であり、アイテム A、B、C、D からなる。シーケンスデータはアイテムとタイムスタンプの組のリストからなる。例えば、ID1 のレコードは、アイテム A、B、C、D がそれぞれタイムスタンプ 1、2、3、4 に現れ、クラス値が Y であることを意味している。

分類パターンマイニングにおける分類モデル構築手順を説明する。ユーザによって支持度(全ユーザ数に対する分類パターンを含むユーザ数の割合)と分類度(ある分類パターン含むユーザ数に対するあるクラスをもつユーザ数の割合)の最小値の条件が与えられる。はじめにデータベース全体から最小支持度を満たす分類パターンを抽出し、分類パターンを前提条件、クラスを結論とするルールを作成し、最小分類度を満たすルールを抽出する。そのうち、支持度の最も高いルールを 1 つ選出し、分類モデルに採用する。採用したルールで判別されるユーザのレコードをデータベースから削除する。そして、残ったレコードからなるデータベースに対して再び 1 つのルールを抽出する処理、分類モデルに追加する処理、追加されたルールに当てはまるユーザのレコードをデータベースから削除する処理を繰り返す。これら処理を、分類パターンが抽出されなくなる、または、データベースのレコードがなくなるまで行い、ルールを採用された順に並べたものを分類モデルとする。

3. 分類パターン

3.1 出現順序を考慮した分類パターン

従来研究では分類パターン抽出の際に、データ種類間の時間軸上での順序を考慮していなかった。しかし、どの種類のデータもタイムスタ

ンプを持つ場合、各種類のデータから抽出されたパターンを、出現順序を考慮して組み合わせることができる。これにより、従来は見つけることのできなかった、より精度の高いルールにより分類モデルの分類精度の向上が期待できる。

本研究では時系列パターンとアイテムセットについて出現順序を考慮し、時系列パターンとその前後に起きたアイテムの組合せを結合した分類パターンを抽出する。

3.2 本研究で抽出する分類パターン

アイテム数 n のアイテム集合 I において k 番目のアイテムを i_k ($i_k \in I, 1 \leqq k \leqq n$)、アイテム i_k のタイムスタンプを t_k とする。

本研究では分類パターンとして、時系列パターン Seq、アイテムセット It、時系列パターンとアイテムセットを出現順序を考慮せずに結合した Seq+It、時系列パターンとその前に起きたアイテムの組合せを繋げた It-Seq、時系列パターンとその後に起きたアイテムの組合せを繋げた Seq-It の5種類を考える。以下にそれぞれの分類パターンを定義する。

Seq：$\{i_{s1}\text{-}\ldots\text{-}i_{sa}\}$　　$(t_{si} < t_{s(i+1)})$

It：$\{i_{i1},\ldots,i_{ib}\}$

Seq+It：$\{(i_{s1}\text{-}\ldots\text{-}i_{sa}),i_{i1},..,i_{ib}\}$

It-Seq：$\{(i_{i1},\cdots,i_{ib})\text{-}i_{s1}\ldots\text{-}i_{sa}\}$
$(t_{ik} < t_{s1},\ 1 \leqq k \leqq b)$

Seq-It：$\{i_{s1}\text{-}\ldots\text{-}i_{sa}\text{-}(i_{i1},\ldots,i_{ib})\}$
$(t_{sa} < t_{ik},\ 1 \leqq k \leqq b)$

ここで、表 1 のデータベースを用いて分類パターンの例を示す。Seq+It の例として(A-B),C,D を説明する。この分類パターンは、時系列パターン A-B を含み、さらにアイテム C と D を含むことを意味する。この分類パターンにあてはまるユーザは ID1、2、3、4、5、6 であり、各クラスのユーザ数はそれぞれ Y が 3、N が 3 のため、(A-B),C,D のときクラス Y のルールの分類精度は 0.50 となる。

It-Seq の例として(C,D)-A-B を説明する。この分類パターンは、時系列パターン A-B を含み、A が最初に現れるタイムスタンプよりも前にアイテム C と D が現れることを意味する。この分類パターンにあてはまるユーザは ID2、3、4 であり、各クラスのユーザ数はそれぞれ Y が 2、N が 1 のため、(C,D)-A-B のときクラス Y のルールの分類精度は 0.67 となる。

Seq-It の例として A-B-(C,D)を説明する。この分類パターンは、時系列パターン A-B を含み、B が最後に現れるタイムスタンプよりも後にアイテム C と D が現れることを意味する。この分類パターンにあてはまるユーザは ID1、2、3、4 であり、各クラスのユーザ数はそれぞれ Y が 3、N が 1 のため、A-B-(C,D)のときクラス Y のルールの分類精度は 0.75 となる。

4. 評価実験

本実験では SNS ゲームログの実データを使用した。このデータはユーザの一回の操作がアイテムとなるシーケンスデータである。クラス値を課金者(Y)と非課金者(N)とし、課金者であるかの分類を行った。データのユーザは 10 万件であり、1 割が課金者である。

ここでは、Seq、It、Seq+It の 3 種類の分類パターンによる分類モデルを従来モデル、Seq、It、Seq+It、It-Seq、Seq-It の5種類の分類パターンによる分類モデルを提案モデルとした。実験データをランダムに 4:1 に分類モデル構築用データと評価用データに分けた。分類モデル構築データから従来モデルと提案モデルをそれぞれ作成し、評価用データに対して各分類モデルを適用して分類を行い、分類精度 A(正しく分類されるユーザ数/分類されるユーザ数)と、カバー率 C(分類されるユーザ数/全ユーザ数)を求めた。

表 2 に分類モデル適用時の評価値と、分類モデル構築時時に得た期待値を示す。

表 2　評価実験の結果

	評価値 (A, C)	期待値(A, C)
提案モデル	0.94, 0.43	0.95, 0.41
従来モデル	0.89, 0.48	0.92, 0.47

実験結果から、出現順序を考慮した分類パターンも含めて抽出することにより分類精度が向上することが確認できた。

6. おわりに

本稿では、時系列パターンとアイテムセットの出現順序を考慮した分類パターンを提案し、この分類パターンによる分類モデルの分類精度が向上することを確認した。

謝辞　本研究は JSPS 科研費 25280022 の助成を受けたものです。

参考文献

[1] B.Liu, W.Hsu, Y.Ma: "Integrating Classification and Assosiation Rule Mining", KDD, 1998.

[2] D.Patel, W.Hsu, M.Lee: "Integrating Frequent Pattern Mining from Multiple Data Domains for Classification.", IEEE ICDE, pp.1001-1012, 2012.