2G3-5in

実数値空間上の頻出パターン最大化によるパターン抽出法

Frequent Pattern Mining by Maximizing Frequent Pattern in Numerical Space

稲場　大樹[*1*2]　福井　健一[*2]　佐藤　一永[*3]　水崎　純一郎[*4]　沼尾　正行[*2]
Daiki Inaba　Ken-ichi Fukui　Kazuhisa Sato　Junichiro Mizusaki　Masayuki Numao

[*1]大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University

[*2]大阪大学産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

[*3]東北大学工学研究科
Graduate School of Engineering, Tohoku University

[*4]東北大学多元物質科学研究所
Institute of Multidisciplinary Research for Advanced Materials, Tohoku University

Frequent pattern mining is a method which extracts item set appearing frequently in the database. The conventional methods can apply for the symbolic data, however, this paper proposes the frequent pattern mining mothod that can apply for numerical data. In the proposed method, labels are assigned to numerical data based on the result of hierarchical clustering, and frequent patterns are extracted the most in the numerical space at the same time. The property is evaluated by using synthetic data. In addition, the proposed method was applied to the data obtained from the damage evaluation test of a fuel cell in order to validate effectiveness against practical data.

1. はじめに

頻出パターン抽出は Apriori アルゴリズム [1] に代表されるように、複数の事象 (アイテム) から成る集合の中で、ある一定以上の頻度で現れるアイテムの組み合わせを列挙する問題である。対象は主に記号で表現されるアイテム集合であるが、QLIQUE[2] などのような数値アイテムに対しても頻出パターン抽出を行う手法が提案されている [3]。この手法は、まず部分空間クラスタリングによって数値アイテムを記号化し、パターン抽出を行う手法である。数値アイテムは< (属性) : (区間値) >という形式で表されるように、1 つのアイテムが 1 つの属性と 1 つの区間値で表される。その一方、数値で表されるデータの中でも座標や波形などといったデータ [4] は、1 つの事象が複数の数値によって記述される。そのような実数値空間上のデータは我々の実世界に数多く存在する。

実数値空間上のデータにおいて頻出パターン抽出を行う場合も、数値アイテムの場合と同様にクラスタリングによって記号化し、パターン抽出を行う手法が考えられる。しかしながら、このような手法はクラスタリングとパターン抽出が両者別々に行われている。そのため、クラスタリングにおいてパターン抽出は考慮されておらず、結果として実際はパターンに関係しない要素もクラスタに含まれてしまい、適切にパターンが抽出されない可能性がある。

そこで、本研究では実数値空間上においても適切にパターン抽出が行われるような、新たな手法を提案する。本手法は、抽出されるパターンを考慮しながらクラスタリングを行い、パターン抽出を行う手法である。

次節では、本手法の概要を説明する。第 3 節では本手法の性能評価実験の結果を示し、さらに第 4 節では燃料電池の損傷評価試験データへの適用結果について述べる。

連絡先: 稲場　大樹，大阪大学産業科学研究所，
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1，
Tel:06-6879-8426，Fax:06-6879-8428，
E-mail:d-inaba@ai.sanken.osaka-u.ac.jp

2. 提案手法：頻出パターン最大化によるパターン抽出

2.1 アルゴリズム

まず、n 個のデータ x_i $(i = 1, \cdots, n)$ 間の距離（非類似度）$D(x_i, x_j)$ が計算できるデータセットが与えられているとする。また、データセットはバスケットに分割された集合として得られているとする。

本手法は、パターンが最も多く抽出できるよう、階層型クラスタリングにおける結合過程、すなわちデンドログラムを基にして適応的に実数値空間上のデータに記号ラベルを付加し、パターンを抽出する手法である。そして、抽出される頻出パターンが最大となったときの頻出パターンを出力する。対象として、n 次元の実数値空間上において、2 つのデータ（事象）間の距離が与えられるようなデータを対象としている。

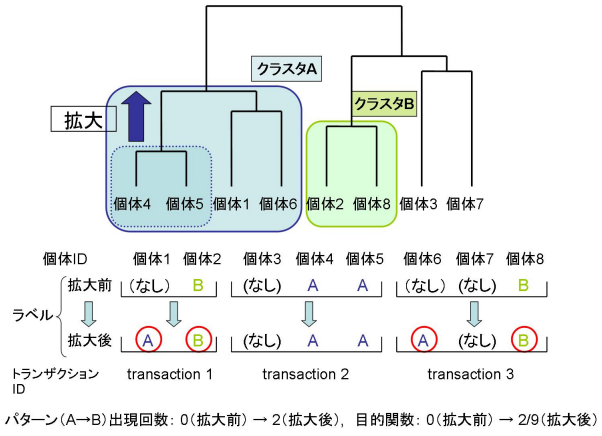

図 1: 提案手法のアルゴリズム模式図

図 1 は、本手法におけるクラスタ拡大とラベル付けの例を、模式図として示したものである。図 1 においてクラスタ A を拡大することで、新たに個体 1 と個体 6 にラベル A が与えられる。このときトランザクション 1 と 3 で A と B が共に現れ、つまりパターン（A → B）が 2 つ現れる。

従来の手法では、クラスタリングによるラベル付けと頻出パターン抽出が別々に行われている。そのため、ラベル付けにおいてパターンが考慮されず、実際にはパターンに寄与しない要素にもラベルを与える可能性がある。したがって、抽出されたパターンのうち、実際に頻出パターンとして存在しているパターンであるかどうか判断するのが困難である。それに対して、本手法ではクラスタリングによるラベル付けと頻出パターン抽出を同時に行っている。クラスタを拡大して記号ラベルを与える際に、後述の 2.2 節に示すようにクラスタにおける個体同士の密集度と確信度を最大化するような頻出パターンの目的関数を用い、これを最大化するように拡大している。そのため、図 1 においてはクラスタを拡大することで目的関数値が改善されるため、クラスタ拡大後において抽出されるパターンの方が、クラスタ拡大前よりも適切なパターンとされる。従って、パターンに寄与する要素にラベルを与えることができ、実際に頻出パターンとして存在しているパターンを適切に、かつ出来るだけ多く抽出することが出来る。

以下に本手法のアルゴリズムを示す。前提条件として、階層型クラスタリングによってデンドログラムが得られているとする。

頻出パターン最大化アルゴリズム

1. n 個の個体の 1 つ 1 つを $Seed_1, Seed_2, \ldots, Seed_n$ とする。そのタネの中から任意の 2 つのタネをそれぞれ $Seed_A$、$Seed_B$ として選択し、それぞれをクラスタ A、クラスタ B とする。$L(A \Rightarrow B) = 0$ に初期化する。
2. もし全ての $Seed$ の組み合わせを探索したならば終了。そうでないならステップ 3 へ。
3. クラスタ A の個体、クラスタ B の個体にそれぞれラベル A、ラベル B をつける。さらに、目的関数値 L を後述の式（1）により計算し、以前の目的関数値より大きいならば L を更新する。
4. $A \cap B = \emptyset$ を満たす全ての組み合わせを探索したら、ステップ 6 へ。そうでないならステップ 5 へ。
5. クラスタ A(クラスタ B) の階層を 1 段階上げてクラスタを拡大する。もし $A \cap B \neq \emptyset$ となるならばクラスタ B（クラスタ A）の階層を 1 段階上げてクラスタを拡大し、クラスタ A（クラスタ B）の階層を $Seed_A(Seed_B)$ と同じ階層に戻す。拡大したクラスタをそれぞれ新たにクラスタ A、クラスタ B とする。ステップ 3 へ。
6. 最小目的関数値、支持度、最小支持度をそれぞれ L_{min}、Sup、Sup_{min} として、$L > L_{min}$ かつ $Sup > Sup_{min}$ ならば頻出パターン $A \Rightarrow B$ として出力。ステップ 1 へ。

2.2 目的関数

頻出パターンの最大化における目的関数の候補として、確信度が考えられる。確信度を上げることは、事象 A が起こったときに事象 B が起こる確率を上げるということを意味する。ただし、ここで問題となるのはクラスタ A、B 双方が拡大するほど事象 A と B が共起する数が大きくなるが、反対にクラスタの密集度が低下して類似性の低い個体を含んでしまう。そこで、本手法では頻出パターンとクラスタの密度の両方を最大化させるために、以下に示すような目的関数 $L(A \Rightarrow B)$ を定義して用いることにする。

$$\begin{aligned} L(A \Rightarrow B) &= \frac{confidence(A \Rightarrow B)}{W(A,B)} \\ &= \frac{1}{W(A,B)} \frac{count(A \cap B)}{count(A)} \end{aligned} \quad (1)$$

ここで、$count(A)$ は事象 A を含むトランザクション数である。そのため、1 つのトランザクションで事象 A が複数回出現しても $count(A)$ では 1 回分としてカウントされる。

また、$W(A,B)$ はクラスタ内の個体の密集度に基づいた重みであり、

$$\begin{aligned} W(A,B) = &(Seed_A\text{からクラスタ } A \text{ を拡大した回数}) \\ &+ (Seed_B\text{からクラスタ } B \text{ を拡大した回数}) + 1 \end{aligned} \quad (2)$$

である。この式 (1) はクラスタが拡大するほど、すなわちデンドログラムにおいて階層が上がるほど類似性の低い個体が含まれ、クラスタ内の個体の密集度が低くなると仮定し、それに応じて変化するような目的関数である。

抽出されたパターンにおいてある値を閾値として、目的関数値がこの値より大きければ相関性の高いパターンと見なして出力する。そこで、この閾値を最小目的関数値として定義する。ただし、式 (1) において目的関数値が高くてもパターン出現回数が少ないパターンも抽出される可能性がある。そこで、本手法ではさらに最小支持度を閾値として用い、抽出されたパターンにおいて出現回数が少ない、すなわち、支持度が最小支持度未満ならば出力しないようにする。

3. 評価実験

3.1 人工データ

まず、人工データを用いて提案手法の特性を評価した。用いた人工データについて、図 2（右図）に示すように、まずは頻出パターン（A → B）としてあらかじめバスケットに正解クラス A および B を割り当てる。図 2 は支持度 0.6、確信度 0.9 の場合である。次に、2 つの正規分布を発生させ、パターンに関係する個体として、その正規分布の中心からの距離が小さい順に個体を選択し、先ほどの A、B それぞれに割り当てる。最後に、残ったデータをノイズとしてランダムにバスケットに割り当てた。全データ数 1000 に対して、各クラスにおける個体数は 500、トランザクション数は 500 とした。

3.2 目的関数値の変化

人工データに対するデンドログラムに基づいた探索における目的関数の値の変化を図 3 に示す。図 3（左図）はクラスタ B の個数、右図はクラスタ A の個数が固定である。クラスタ内の個体数、すなわちクラスタが小さい場合、クラスタが拡大するにつれて事象 A と B が共起する確率が高くなるため、目的関数値が増加する傾向にある。しかし、ある程度クラスタが

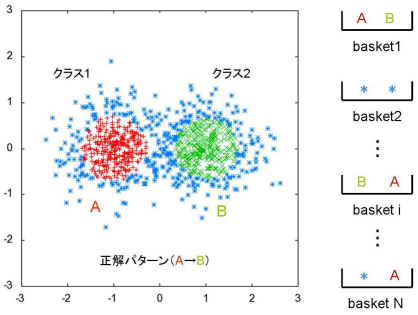

図 2: 人工データとバスケット生成

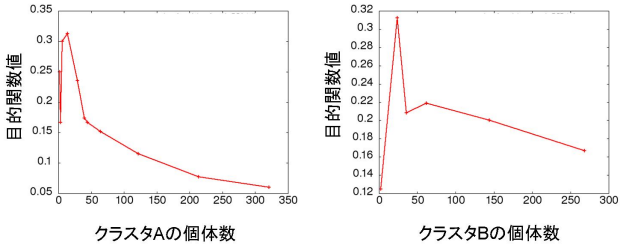

図 3: クラスタ内の個体数に対する目的関数値の変化

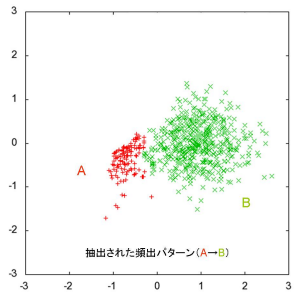

図 4: 頻出パターンの抽出結果

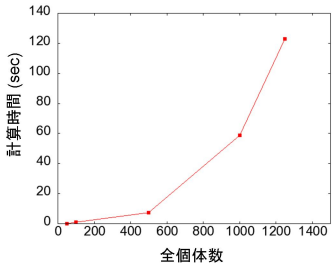

図 5: 全個体数に対する計算時間

拡大すると事象 A と B が共起する確率が高くなるより、クラスタ内の個体の密集度が低下する傾向が強まり、目的関数値が減少する傾向にあることが分かる。抽出されるパターンは目的関数値が最大のときである。したがって、目的関数として式 (1) を用いることで、パターンの出現回数とクラスタの密集度の両方が最大化されていることが分かる。

3.3 抽出精度

次に、パターン抽出精度について表 1 に示す。表 1 において正解パターンの確信度は 0.9 であり、抽出された頻出パターンのうち最も目的関数値が大きいパターンについて評価している。評価尺度として、精度、再現率、F 値を用いた。これらの値がそれぞれ、大きければ大きいほどパターンの抽出精度が高いと見ることが出来る。

表 1 の結果において、支持度が低くなるほど精度が低下し、再現率が増加している。これは、正解パターンに対してノイズの影響が大きくなるため、パターンに関係する部分のクラスタを得ることが困難であるからと考えられる。それに対し、支持度が大きい場合では精度が高く再現率が低い。これは、デンドログラムにおいて階層の上位まで上げなければならないため、目的関数におけるクラスタ内の個体の密集度の影響を受けて、パターンに関係する部分より小さなクラスタが生成されたからであると考えられる。しかしながら、F 値で見ると支持度によって値に大きな変化がないため、パターンの出現回数によらず頻出パターンを抽出できる点は本手法の特徴であると言える。

表 1: 正解パターンにおける各支持度に対する精度 (Precision)、再現率 (Recall)、F 値

支持度	Precision	Recall	F 値
0.8	0.842	0.737	0.786
0.7	0.798	0.772	0.785
0.6	0.702	0.820	0.756

また、図 4 は支持度 0.6、確信度 0.9 において実際に抽出された頻出パターンである。クラスタリングと頻出パターン抽出を別々に行う従来手法では、図 4 のように分布の一部をクラスタとして得ることは困難である。それに対し、本手法ではクラスタ A として正解パターンの分布の約半分、クラスタ B としてクラス 2 の分布のほぼ全てをパターンとして抽出されている。しかしながら、本手法はデンドログラムに基づいて探索を行っているため、生成されるクラスタの形状がデンドログラムに依存した形状になる。そのため、非類似度を計算する際にもパターンを考慮する必要がある。

3.4 計算時間

計算時間について図 5 に示す。実験には、CPU : Intel Xeon Quad-core 3.2GHz 、RAM : 16GB のコンピュータを用いた。この結果によると、データ数が一桁増加するにつれて、計算時間が指数関数的に増大していることが分かる。探索回数は理論上、個体数 n 個に対してクラスタの $Seed$ の組み合わせが n^2 回、さらにそれぞれがデンドログラムの階層の深さ $\log n$ 回の探索を行うため、全体で $(n \log n)^2$ 回の探索が行われる。したがって、計算時間が大きく、今後計算時間の短縮のため改良が望まれる。

4. 実データへの適用－燃料電池の損傷評価

4.1 損傷評価試験の概要

次に、実データへの応用例として、燃料電池の損傷評価試験から得られたデータに本手法を用い、頻出である損傷パターンの抽出を行った。燃料電池は高温・酸化還元である厳しい環境で稼動しており、亀裂・はく離などの物理的損傷が生じる [5]。

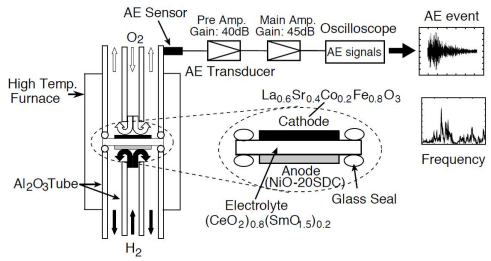

図 6: SOFC 損傷試験装置

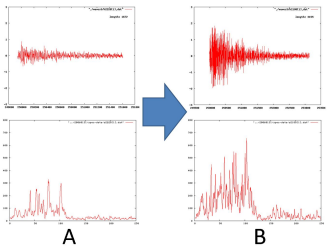

図 7: 抽出された損傷パターンの例（上段：AE 波形、下段：AE の周波数スペクトル）

この際、AE(Acoustic Emission) と呼ばれる微弱な弾性波（すなわち振動、超音波）が発生し、この波形を分析することで、損傷の種類や損傷が発生した部材を特定できる。図 6 はその測定装置の概略図である。

実際、この AE 事象は大量かつ様々な種類に対応した波形が発生しており、本手法を用いてこれらの AE 事象から頻出である AE 事象の組み合わせを探す。これにより、燃料電池における構成部材間の力学的な相関関係を知ることが出来る。

本手法を適用する前に、上記の AE 事象は一連の事象の時系列として存在している。そこで、従来研究 [6] に基づき、ある一定値のエネルギーを超えたときに蓄積された力が解放されると仮定し、それまでを一連の損傷とみなし、AE 事象の列をバスケットに分割した。分割した 1 つの区間を 1 つのバスケットとして扱い、この短い区間での AE 事象の共起性を調べる。本研究では、実験装置を 60 時間稼動させて得られた 1429 個の AE 事象を用いた。これらの AE 事象は上記の方法で 148 個の区間に分割された。AE 事象間の距離としては、離散周波数スペクトルに対して先行研究 [5] において高い分類性能を示した、Kullback-Leibler 情報量に基づく距離を用いた。

4.2 損傷パターン抽出結果

パターン抽出におけるパラメータとして、最小支持度 $Sup_{min} = 0.025$、最小目的関数値 $L_{min} = 0.30$ において 9 種類の損傷パターンが抽出された。図 7 はその一例である。この損傷パターンにおける支持度は 0.027、すなわちパターン出現回数は 4 回であり、確信度は 0.8 であった。この損傷パターンは構成部材がきしむことで、同時に初期欠陥や材料の不均一性から生じた割れの進展を引き起こすといったパターンである。

抽出された 9 種類の損傷パターンについて、燃料電池の専門家の評価によれば、燃料電池における構成部材間の潜在的な力学的相関関係を示す非常に興味深い結果であることが分かった。

5. 今後の展望

本手法は実数値空間における頻出パターン抽出法として、燃料電池の損傷評価試験から得られる AE だけでなく、地震波などの波形データや、動画像における動点軌跡などの様々な実数値空間上のデータに適用可能であることが期待できる。また、自己組織化マップにおける参照ベクトルを 1 つの個体と見なし、デンドログラムを作成することで、本手法と同様にして頻出パターンが得られると思われる。これにより、パターンの全体像を直感的に把握することを容易にできると期待される。

6. まとめ

本論文では、実数値空間上における新たな頻出パターン抽出法を提案した。従来までの実数値空間上におけるパターン抽出法では、クラスタリングによる記号化とパターン抽出が別々に行われていた。それに対して本手法では、抽出されるパターンとクラスタの密集度が最大となるように、デンドログラムに基づき適応的にクラスタを生成し、パターン抽出を行う手法である。この手法は、従来手法よりも適切にパターンを抽出することが出来る。また、燃料電池の損傷評価に応用し、実データに対しても有効であることを示した。

謝辞

本研究は「ナノマクロ物質・デバイス・システム創製アライアンス」特別経費（文部科学省）、および科学研究費補助金若手研究 (B)(21700165) の支援を受けて行われた。

参考文献

[1] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules”, *Proc. of 20th Int. Conf. on Very Large Databases*, pp.487-499, 1994.

[2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, “Automatic subspace clustering of high dimensional data for data mining applications”, *Proc. of the 1998 ACM SIGKDD international conference on Management of data*, pp.94-105, 1998.

[3] 光永悠紀, 鷲尾隆, 元田浩, “適応的密度基準に基づく部分空間クラスタリングを用いた定量的多頻度アイテム集合のマイニング”, 人工知能学会論文誌, Vol.21, No.5, pp.439-449, 2006.

[4] B. Samanta and C. Nataraj, “Use of particle swarm optimization for machinery fault detection”, *Engineer Application of Artificial Intelligence*, vol.22, pp.308-316, 2009.

[5] 福井健一, 赤崎省悟, 佐藤一永, 水崎純一郎, 森山甲一, 栗原聡, 沼尾正行, “固体酸化物燃料電池における損傷過程の可視化”, 日本機械学会論文集 A 編, Vol.76, No.762, pp.223-232, 2010.

[6] 北川哲平, 福井健一, 佐藤一永, 水崎純一郎, 沼尾正行, ”“キーグラフと SOM を用いた稀な重要事象抽出による燃料電池の損傷評価”, 情報処理学会論文誌:数理モデル化と応用, Vol.4, No.2, pp.1-12, 2011.