■7 群（コンピュータソフトウェア）－ 6 編（情報検索とデータマイニング）

6 章 データマイニングの基礎技術

【本章の構成】

本章では以下について解説する．

6-1 頻出パターンの抽出

6-2 リンク解析と可視化

6-3 嗜好学習とリコメンデーション

6-4 時系列解析，パターン発見，変化点検出

6-5 外れ値検出・不正検出

■7 群－6 編－6 章

6-1 頻出パターンの抽出

（執筆者：森本康彦）［2009 年 1 月 受領］

データベースから頻出するパターンを抽出する問題は，「データマイニング」という用語が定着するきっかけとなった相関ルール発見問題の部分問題として広く認知され応用されてきたが，データ中に現れる頻出パターン自体がデータの特徴を捉える重要な知見となりえるため，当初より情報検索やデータマイニングにおける最も重要な基礎技術として盛んに研究されてきた．現在ではデータマイニング技術の応用分野は多岐にわたっているため，データ中の「頻出パターン」たる部分構造には様々なものがあるが，本節では，トランザクションデータベースから頻出アイテム集合を抽出する問題を仮定して解説を進める．しかしながら，その一般性は高く，トランザクションデータベース以外で利用されている頻出パターン抽出技術の多くも，広い意味ではこの頻出アイテム集合抽出技術に基づいている．

6-1-1 頻出アイテム集合

いま，**表 1･1** のようなトランザクションデータベースがあるとする．表の各行はトランザクションの内容を表しており，例えば，100 番のトランザクションではアイテム A, C, D が，同様に 200 番ではアイテム B, C, E が購入されたことを表している．X をデータベース中のあるアイテム集合とする．X を含むトランザクション数を X の支持度と呼び，本節ではこれを sup(X) と表す．頻出アイテム集合とは，ユーザが指定する最小支持度以上の頻度でデータベースに存在するアイテム集合のことを指す．

2 つのアイテム集合 X, Y が X⊆Y である場合，その支持度には sup(X)≥sup(Y)となる「逆単調性」と呼ばれる性質がある．例えば，アイテム集合{B, C}では sup({B, C})=2 であるが，これを部分集合として含む，{A, B, C}などのいかなるアイテム集合もその支持度は 2 以下となる．代表的な頻出アイテム集合抽出アルゴリズムはこの逆単調性を利用して効率的に頻出アイテム集合を求める．

表 1・1

TID	Item
100	A, C, D
200	B, C, E
300	A, B, C, E
400	B, E

6-1-2 アプリオリ

データベース中に存在するアイテム数を m とすると，アイテム集合は空集合を除いて全部で 2^m-1 個ある．表 1･1 のデータベース中に存在する 5 つのアイテム集合を束で表したものが **図 1･1** である．束の枝は各集合の包含関係を表しているが，前述の逆単調性は，この束では，上位頂点の支持度は，その下位頂点の支持度以下であるということを示している．文献 1) のアプリオリと呼ばれるアルゴリズムは，この束を広さ優先探索する．アプリオリでは，ある頂

点が頻出ではないと判明したら，その頂点より上位にある頂点の探索をしないことで効率化を図っている．

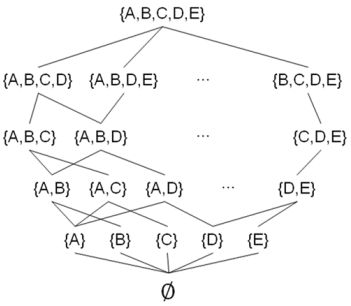

図 1・1　アイテム集合の束

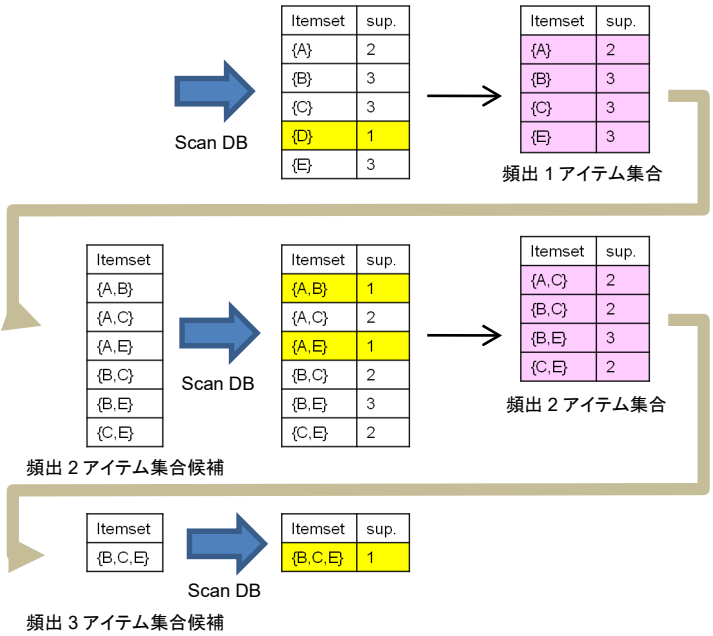

図 1・2　アプリオリ探索

図 1･2 にアプリオリアルゴリズムにそって表 1･1 のデータベースから最小支持度 2 の頻出アイテム集合を列挙していく過程を示す．ここで要素数 k のアイテム集合を k アイテム集合とする．最初にデータベースを走査し，頻出 1 アイテム集合を求める．この例では，{D}は支持度が 1 であるため，以降は{D}を含むアイテム集合は探索の対象外となる．頻出 1 アイテム集合

の組合せで作られる 2 アイテム集合を作り，それを頻出 2 アイテム集合候補とする．候補を求めたら，それぞれの候補に対する支持度をデータベース走査により求め，頻出 2 アイテム集合を求める．次に頻出 2 アイテム集合から頻出 3 アイテム集合候補を作るが，候補となりうる 3 アイテム集合はその部分集合がすべて頻出でなくてはならない．このデータベースの場合，例えば{A, B, C}は{A, C}と{B, C}は頻出であるが{A, B}が頻出でないため逆単調性から候補とはならない．このようにして頻出 k アイテム集合から $k+1$ アイテム集合の候補を作り，データベース走査により頻出 $k+1$ アイテム集合を求めるという操作を繰り返していく．

6-1-3 高速化・大規模化

アプリオリアルゴリズムが発表された以降も，データベースの更なる大容量化や，高速計算への要求は高まり，その後も様々な頻出パターン抽出技術が開発されている．

アプリオリアルゴリズムは，その他の頻出アイテム集合抽出アルゴリズムに比較して，データベースの走査回数が少ないことと，データベースのサイズがメモリに保持しきれないくらい大規模であっても高速に動作することが特徴としてあげられるが，頻出アイテム集合をすべてメモリに保持する必要があり，これが問題となることもある．

文献 4)では，データベース内に存在するアイテム集合の支持度などの基本情報を，空間効率の良い頻出パターン木と呼ばれるデータ構造で維持管理する手法でアプリオリにおけるメモリ問題を解決している．

文献 3)のバックトラック法は，アプリオリとは対照的に，束を深さ優先で探索していくアルゴリズムである．バックトラック法は，アイテムごとの（垂直配置の）トランザクション情報がすべてメモリに収まる場合には，アプリオリより高速であることがいくつかの比較実験により示されている．

頻出パターン抽出技術に関するサーベイとしては文献 5)が詳しく，頻出パターン木や，バックトラック法，その他の技術の詳細に関してはこの論文を参考とするのがよい．

■参考文献

1) R. Agrawal and R. Srikant：“Fast Algorithms for Mining Association Rules in Large Databases,” Proc. of VLDB Conference, pp. pp.487-499, 1994.
2) R. Agrawal, T. Imielinski, and A. Swami：“Mining association rules between sets of items in large databases,” Proc. of ACM SIGMOD Conference, pp. 207-216, 1993.
3) R. Bayardo：“Efficiently mining long patterns from databases,” Proc. of ACM SIGMOD Conference, pp.85-93, 1998.
4) J. Han, J. Pei, and Y. Yin：“Mining frequent patterns without candidate generation,” Proc. of ACM SIGMOD Conference, pp.1-12, 2000.
5) 宇野毅明，有村博紀：“データインテンシブコンピューティングーその 2 頻出アイテム集合発見アルゴリズムー,” 人工知能学会誌, vol.22, no.3, pp.425-436, 2007.
6) 加藤直樹，羽室行信，矢田勝俊：データマイニングとその応用，朝倉書店, 2008.
7) 福田剛志，森本康彦，徳山 豪：データマイニング，共立出版, 2001.

■7 群－6 編－6 章

6-2 リンク解析と可視化

（執筆者：藤村　考）［2009 年 6 月 受領］

リンク解析（Link Analysis）は，データマイニングの一分野であり，マイニング対象のデータをノード V とエッジ E から構成されるグラフ $G = (V, E)$ として捉え，グラフ G の構造から有用な知識を発見しようとする技術である．

リンク解析を行うにあたって先ず考えなくてはならないことは，分析対象のデータから何をノードとし，何をエッジとして抽出するかということと，リンク解析のアルゴリズムとして何を適用するかということである．これらのポイントは，マイニング対象のデータからどのような知識を発見したいのかに依存する．リンク解析は，その応用が広がっており，すべてのトピックを網羅することはできないので，本節では，代表的なリンク解析アルゴリズムである PageRank [1] と HITS [2] について説明するとともに，その応用を述べる．

なお，リンク解析と呼ぶ技術には，これらの手法だけではなく，G のトポロジーからクラスタに分割する手法や k-クリーク，k-コア，k-デンスなどのサブネットワークを抽出する手法など多数存在する．これらに関しては参考文献 3)にコンパクトにまとめられている．また，リンク分析の近年の話題としては，3 ノードから構成される局所ネットワークの特定のパターンが G にどれくらいの確率分布で含まれるかを分析するモチーフ分析 [4] と呼ばれる手法も盛んに行われるようになっている．

また，グラフ G を視覚的にわかりやすく可視化することは，リンク解析の対象を事前に把握したり，リンク解析の結果を直観的に理解するために有効であることから，本節ではグラフの可視化技術についても併せて説明する．

6-2-1 PageRank

PageRank [1] は，Web 空間から Web ページをノード，Web ページ間のハイパーリンクをエッジとみなす大規模なグラフとみなし，その構造からノードの重要性をランキングするものであり，次式で表現される．

$$\mathbf{p} = [\alpha\mathbf{H} + (1-\alpha)\mathbf{A}^{\mathrm{T}}]\mathbf{p} \qquad (2 \cdot 1)$$

ただし，α は $[0, 1]$ を範囲とする定数，$\mathbf{H}$ は任意の要素 $h_{ij} = 1/n$（n は解析対象の全ノードの要素数）とする n 行 n 列の行列である．$\mathbf{A}$ は，ノード i から j へのリンクがある場合 $a_{ij} = 1/n_i$，ない場合は $a_{ij} = 0$ を要素とする隣接行列である．ここで n_i はノード i のリンクの出次数である．$\mathbf{p}$ は各ノード i のページランクスコア p_i を要素とするベクトルである．

上式において，$\mathbf{p} = \mathbf{A}^{\mathrm{T}}\mathbf{p}$ の項は「重要なページからリンクされているページは重要である」という再帰的な関係を表現しており，ノード j のスコア p_j は，ノード j のすべてのリンク元のノード i のスコア p_i をノード i の出次数 n_i で正規化した値を集計したものであり，$\mathbf{p} = \mathbf{Hp}$ の項は，ランダムジャンプと呼ばれ，すべてのノードのスコア p_i をノード数 n で正規化した値の集計である．これら 2 つの項が重み α で線形結合されている．

PageRank の算出は，$\mathbf{M} = \alpha\mathbf{H} + (1-\alpha)\mathbf{A}$ とするとき，$\mathbf{p}$ に $[1, \cdots, 1]^{\mathrm{T}}$ の初期ベクトルを与え，$\mathbf{M}^{\mathrm{T}}$ を繰り返し乗じることで算出できる．

6-2-2 HITS

HITS（Hypertext Induced Topic Selection）[2] は，PageRank と同様に，ハイパーリンクで結合された Web ページをグラフとするネットワーク構造から，主要なトピックに関連するサブグラフを抽出するアルゴリズムであり，次式で表現される．

$$\mathbf{a} = \mathbf{A}^{\mathrm{T}}\mathbf{h} \qquad (2・2)$$

$$\mathbf{h} = \mathbf{A}\ \mathbf{a} \qquad (2・3)$$

ただし，$\mathbf{A}$ は，ノード i から j へのリンクがある場合 $a_{ij}=1$，ない場合は $a_{ij}=0$ を要素とするハイパーリンクを表現する隣接行列である．$\mathbf{a}$ は各ノード i のオーソリティスコア a_i を要素とするベクトルである．$\mathbf{h}$ は各ノード i のハブスコア h_i を要素とするベクトルである．

上式において，$\mathbf{a} = \mathbf{A}^{\mathrm{T}}\mathbf{h}$ は「重要なハブとなるページからリンクされているページはオーソリティである」という関係を，$\mathbf{h}=\mathbf{A}\mathbf{a}$ は「重要なオーソリティとなるページをリンクしているページは重要なハブページである」を表現し，ノード j のオーソリティスコアは，リンク元のノード i のハブスコア h_i の集計であり，ノード i のハブスコアは，リンク先のノード j のオーソリティスコア a_j の集計である．

HITS は，上式を変形し $\mathbf{a}=(\mathbf{A}^{\mathrm{T}}\mathbf{A})\mathbf{a}$ としたとき，$\mathbf{a}$ に関する再帰式となる．そこで，オーソリティスコアベクトル $\mathbf{a}$ は，$[1, \cdots, 1]^{\mathrm{T}}$ の初期ベクトルを与え，$\mathbf{A}^{\mathrm{T}}\mathbf{A}$ を繰り返し乗じることで算出できる．ただし，PageRank とは異なり，$\mathbf{A}^{\mathrm{T}}\mathbf{A}$ は正規化されていないので，値が発散しないように，繰り返しごとに $\mathbf{a}=\mathbf{a}/\|\mathbf{a}\|$ を適用し正規化する．

HITS は PageRank とは異なり，ランダムジャンプの項がないため，繰り返しアルゴリズムにより算出とすると，$\mathbf{a}$ と $\mathbf{h}$ の相互強化により，主要なトピックに属するノード群のスコアだけが極端に高くなり，それ以外のノードスコアは限りなく 0 に近づく．この結果，主要なトピックに所属するノード群が抽出される．したがって，HITS は，PageRank のようにグローバルな Web 空間に適用するのではなく，クエリ q を指定した検索結果の Web ページ集合をノードとするサブグラフ $G_q=(V_q, E_q)$ に対して適用し，その中で主要なトピックに関連する検索結果のみに絞り込む目的で通常使用される．ただし，繰り返しアルゴリズムではなく，$\mathbf{A}^{\mathrm{T}}\mathbf{A}$ の固有ベクトルを算出し，各固有値に対応するオーソリティスコアあるいはハブスコアを算出し，それぞれの固有値に対応するノード群を抽出することにより，トピックごとにクラスタリングする方法として利用することもできる．

6-2-3 PageRank と HITS の応用

PageRank と HITS は，元々 Web ページをノードとし，ハイパーリンクをエッジとするネットワークに適用して，Web の検索に応用することを目的としていたが，近年，その適用先が Web ページ以外の様々な対象に広がっている．例えば，SNS やブログ空間におけるユーザをノードとし，コミュニケーションをエッジとして，PageRank を適用し，コミュニティの中心となるユーザを発見したり，HITS を適用し，(サブ)コミュニティから主要なクラスタを抽出したりするなどの応用である．

また，HITS はハブスコアとオーソリティスコアをそれぞれ異なるドメインの要素に適用することもできる．例えば，ハブスコアの対象となるノードとして「ユーザ」を，オーソリティスコアの対象となるノードとして「商品」としたとき，ユーザ集合と商品集合から構成される二部グラフとなるが，このような二部グラフは，マイニングではしばしば出現する．このよう

な対象に対しても HITS は適用可能であり，様々なデータに対して分析が行われている．

一方，元々の Web の検索というタスクに関しては，Web 2.0 の潮流により，自動的に形成されるハイパーリンクが急速に増加し，また，SEO を目的として故意にリンクネットワークを構築するスパムサイトの出現により，ハイパーリンクだけでは，重要なページが発見できなくなってきている．近年の商用検索エンジンではリンク解析は重要な Web ページの抽出より，むしろスパムの抽出に重要な役割を果たすようになっている．

6-2-4 グラフの可視化

グラフを可視化とは，グラフ G が与えられたときに，その特徴を把握しやすいように表現する技術であり，一般的には，2 次元もしくは 3 次元のユークリッド空間にノードを「点」として，エッジを点と点をつなぐ「直線」として表現して描画する．

リンク解析においては，マイニング対象のデータからグラフ G を抽出するステップがあるが，この段階でマイニング対象のグラフそのものが想定されたとおり抽出できているかどうか把握するため，しばしばグラフの可視化が行われる．また，リンク解析によって抽出された各ノードの重要性などの属性がマイニングによって想定したとおり抽出しているかどうかを把握するためにもグラフの可視化が行われる．この場合には，マイニングによって抽出されたノードあるいはエッジの属性を着色することで，どのようなノードあるいはエッジが抽出されたかを直感的に表現することができる．以上のように，リンク解析とグラフの可視化は，しばしばセットで行われる．

グラフ可視化の古典的な手法としては，まず Eades [5] のバネモデルがある．これは各エッジに自然長を有するバネの存在を仮定し，バネのエネルギーが最少（極小）となる座標を求めるものである．しかし，グラフのトポロジーやサイズによっては，関連のないノードがたまたま近くに配置されるケースが少なくない．そこでバネモデルの改良版としては，Kamada ら [6] の KK 法がある．KK 法では，G のエッジに対応するバネを仮定するのではなく，各ノードのすべての組合せのエッジに対して，各ノード間の最短パス長に比例する自然長を持つバネを仮定することで，上記の問題を改善している．また，より大規模なグラフに対して計算を可能にする方法としては，Fruchterman ら [7] の FR 法がある．FR 法では，エッジが存在するノード間に引力を与え，更にエッジが存在しないノード間に斥力を与えることで系としてエネルギーが最小（極小）となる座標を求めるものである．FR 法でも KK 法でも繰り返し演算により，エネルギーが低くなる方向に座標を変化させることで計算をするが，FR 法では，繰り返しの各ループで，すべてのノードを同時に座標を移動させる効率的な実装が可能であり，大規模なグラフに向いている．このようなアルゴリズムが実装された可視化ツールも多数開発されているが，広く利用されているものとしては，Cytoscape [8]，Pajek [9]，Graphviz [10] などがある．

グラフの可視化における近年の研究としては，GPGPU [11] を使って並列計算を行うことで，より大規模なグラフに対応できるようにするもの [12),14)]．動的に変化するグラフを扱うもの [13)]．各ノードに文字列などを表示した場合，文字同士がお互いに重ならないように配置するなどの新しいニーズに応えるもの [14)] などがある．

■参考文献

1) Brin, S., Page, L.：“The anatomy of a large-scale hypertextual Web search engine,” Proceedings of the Seventh International World Wide Web Conference, pp.107-117, 1998.
2) Kleinberg, J.M.：“Authoritative sources in a hyperlinked environment,” J. ACM, vol.46, no.5, pp.604-632, 1999.
3) 林 幸雄(編)：ネットワーク科学の道具箱, 近代科学社, pp.119-156, 2007.
4) Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon. U.：“Network Motifs: Simple Building Blocks of Complex Networks,” Science, vol.298, no.5594, pp.824-827, 2002.
5) Eades, P.：“A Heuristic for Graph Drawing,” Congressus Numerantium, vol.42, no.149-160, 1984.
6) Kamada, T. and Kawai, S.：“An algorithm for drawing general undirected graphs,” Information Processing Letters, vol.12, no.31, pp.7-15, 1989.
7) Fruchterman, T.M.J. and Reingold, E.M.：“Graph Drawing by Force-directed Placement, Software － Practice and Experience,” vol.11, no.21, pp.1129-1164, 1991.
8) Cytoscape：http://www.cytoscape.org/
9) Pajek：http://vlado.fmf.uni-lj.si/pub/networks/pajek/
10) Graphviz：http://www.graphviz.org/
11) GPGPU：http://www.gpgpu.org/
12) Frishman, Y. and Tal, A.：“Multi-Level Graph Layout on the GPU,” IEEE Trans. Visualization and Computer Graphics, vol.13, no.6, pp.1310-1317, 2007.
13) Frishman, Y. and Tal, A.：“Online Dynamic Graph Drawing,” IEEE Trans. Visualization and Computer Graphics, vol.14, no.4, pp.727-740, 2008.
14) 松林達史, 山田武士, 藤村 滋, 藤村 考：“固有楕円ポテンシャルを利用したラベル付きグラフ可視化の座標計算,” 情処学論: 数理モデル化と応用, vol.1, no.1, pp.88-101, 2008.

■7 群－6 編－6 章

6-3 嗜好学習とリコメンデーション

（執筆者：土方嘉徳）［2008 年 11 月 受領］

情報検索及びデータマイニングの基本概念とアルゴリズムは，嗜好学習とリコメンデーションに利用されてきた．嗜好学習とは，個人の嗜好や興味を，その個人の計算機やインターネット，更にはセンサの埋め込まれた実世界での行動から学習するものである．学習した嗜好や興味に関する情報を用いれば，そのユーザの気に入りそうな商品や，興味を持ちそうな新聞記事などを提示することができる．このような，個人の嗜好や興味に基づく商品やコンテンツの選択をリコメンデーション（情報推薦）と呼ぶ．

多くの嗜好学習や情報推薦の手法は，情報検索やデータマイニングの手法を基にして発展してきた．ユーザがどのような行動をとった末にどの商品やコンテンツを選択したかという過去の事例から，ユーザの嗜好を学習していく．事例から一般化されたモデルを学習する点においては，データマイニングが行っていることと変わりない．しかし，事例として得られる嗜好が不安定であること，そして予測すべき未来の嗜好も不安定であることが，独自の理論体系を生み出してきた．本節では，嗜好学習及び情報推薦の基本的な考え方と，そのアルゴリズムについて説明する．

6-3-1 嗜好学習と情報推薦とは

インターネットの普及に伴い，インターネット上での商取引も盛んに行われるようになった．インターネット上の商用 Web サイトにおいても，実世界の店舗同様に顧客への声かけや提案は重要な販売促進活動と言える．多くの商用 Web サイトでは，ユーザの過去の閲覧（行動）履歴を用いて，そのユーザの嗜好や興味にあった商品やコンテンツ（以降，「アイテム」と呼ぶ）を提案し，表示している．このような提案・提示を一般に「情報推薦」あるいは「リコメンデーション」と呼ぶ．これを実現するシステムを推薦システムという．

嗜好学習は情報推薦を行うには必要不可欠な技術である．ユーザの嗜好や興味を事例から学習し一般化する手法である．ここでいう事例とは，ユーザの過去の閲覧（行動）履歴とアイテムの選択履歴である．ユーザの嗜好を学習し，それに基づきアイテムの推薦を行うというプロセスは，一見それぞれが独立しているように思える．しかし，情報推薦の手法には，嗜好のモデル化を含んだものも多く，一概に分離できるとは限らない．

6-3-2 情報推薦の歴史

インターネットが普及し始めた 1980 年代後半から，流通する情報が爆発的に増加することを見越し，それらの情報からユーザに適したものを取捨選択する情報フィルタリングの考え方が広まった．情報フィルタリングも情報推薦同様アイテムを取捨選択するのではあるが，それは必ずしもユーザ個人の嗜好・興味のモデルに基づいていなくてもよい．代表的な情報フィルタリングのシステムには 1986 年に MIT のマローン（Malone）らが開発した Information Lens [1] がある．これは，電子メールを，電子メールのメッセージのヘッダ部を有効に活用し，ルールベースで電子メールのフィルタリングを試みた研究である．

1990 年代に入ると，情報検索や人工知能の研究者が競って，情報推薦の研究を始めた．代表

的な研究としては，フォルツ（Foltz）らが行った適合性フィードバックと明示的なキーワード入力が文書推薦に与える影響に関する調査 [2] と複数のユーザのアイテムに対する評価履歴を用いた協調フィルタリングの提案 [3] である．これ以降，情報推薦は，情報検索やデータマイニング，機械学習の実験場のようになり，多くのアルゴリズムが提案されてきた．情報推薦は，コンピュータ科学全体から見ても注目を集め，1992 年，1997 年，2000 年に，情報推薦及びパーソナライゼーションの特集号が組まれている [4)~6)].

6-3-3 情報推薦の基本方式

情報推薦の基本方式には，(1) コンテンツに基づくフィルタリング（content-based filtering）と，(2) 協調フィルタリング（collaborative filtering）の 2 種類がある．前者は，推薦するアイテムの内容に基づき情報の取捨選択を行う．後者は，ネットワーク上に存在する同じ好みを持ったコミュニティを発見し，そのコミュニティが共通して好むアイテムを選択する．いずれの方法も，ユーザの過去の行動履歴を基に推薦を行う．そのためには，過去の行動履歴からユーザの嗜好や興味に関する情報を獲得し，それをモデル化（一般化）する必要がある．このモデル化した情報をユーザプロファイル（user profile）と呼ぶ．また，嗜好や興味に関する情報を獲得しモデル化することを，嗜好学習（嗜好抽出，ユーザプロファイリングとも呼ばれる）と呼ぶ．嗜好学習の詳細は，文献 7) が詳しい．

(1) コンテンツに基づくフィルタリング

コンテンツに基づくフィルタリングは，従来の情報検索技術の影響を濃く受けている．その基本的な考え方は，情報検索の分野で提案された適合性フィードバック（relevance feedback）[7] にある．適合性フィードバックの定義は，情報検索において検索結果として出力された文書の内容に基づいて，検索質問や検索戦略，検索式を修正することを指す．コンテンツに基づくフィルタリングにおいては，ユーザからの行動履歴を基にユーザプロファイルを変更することになる．

コンテンツに基づくフィルタリングの基本手法について説明する．まず，推薦対象のコンテンツからコンテンツの特徴量を抽出する．コンテンツがテキストの場合は，キーワードの出現頻度などで表される．抽出した特徴量はモデル化される．このモデル化したものをコンテンツモデルと呼ぶ．ユーザからも，そのコンテンツに対する評価やアンケートなどから，コンテンツの特徴量に関する嗜好情報を抽出し，モデル化する．これがユーザプロファイルとなる．推薦は，コンテンツモデルとユーザプロファイルを比較することで行われる．

実際には，コンテンツやユーザの情報のモデル化は，推薦方式と一体となっていることが多い．推薦方式は大きくは，(1) ルールベース方式（rule-based method），(2) メモリベース方式（memory-based method），(3) モデルベース方式（model-based method）の 3 種類に分けられる．ルールベース方式は，事前に推薦の仕組みをルールで記述しておくものである．メモリベース方式は，コンテンツモデルとユーザプロファイルの両方をベクトルで表し，ベクトル空間上での距離により，推薦するか否かを決定する方式である．モデルベース方式は，過去に閲覧/購読したアイテムに対する評価値から，一般的な興味の傾向をモデル化し，ユーザプロファイルとする方式である．新たなアイテムが発生すれば，それをベクトル形式などでモデル化し，上記ユーザプロファイルと比較する．閲覧/購読を機械学習の教師信号と考え，アイテム（いくつかの特徴量を有する）とそれに対する正負の判断という組を，機械学習のアルゴリズムに入力し

て学習することで得られる．

(2) 協調フィルタリング

協調フィルタリングは，CSCW の影響を受けていると思われる．その基本的な考え方は，口コミ（word of mouth）にある．すなわち，自分と近い興味の人から，お勧めのアイテムを推薦されれば，それを購入/閲覧してしまうというものである．協調フィルタリングでは，自分に近い嗜好や興味を持つユーザ群を発見し，そのユーザ群が高く評価したアイテムで，まだ自分が未評価のアイテムを推薦する．以下に，レズニック（Resnick）らが提案した協調フィルタリングのアルゴリズムを示す．

(a) 協調フィルタリングのアルゴリズム

ユーザ集合を $A=\{a_1,a_2,\cdots,a_n\}$，アイテム集合を $B=\{b_1,b_2,\cdots,b_m\}$ とし，ユーザ a_i がアイテム b_k につけた評価値を $r_{i,k}$ とする．まずは，近傍を形成する．a_i を注目ユーザとしたとき，すべての $a_o \in A$（ただし，a_i を除く）に対する類似度 $s(a_i,a_o)$ が，r_i と r_o の類似度に基づいて計算される．一般には，類似度の計算にピアソン相関やコサイン距離が用いられる．最も似ているユーザ上位 M 人が a_i の近傍メンバーになり，その集合を $neighbor(a_i) \subseteq A$ と表す．

次に，アイテムの評価値の予測を行う．$a_o \in neighbor(a_i)$ が評価をつけており，かつ a_i が未評価であるアイテム b_k すべてに対して，嗜好の予測値 $p_i(b_k)$ が計算される．

$$p_i(b_k)=\bar{r}_i+\frac{\sum_{a_o\in A_i'} s(a_i,a_o)\cdot(r_{o,k}-\bar{r})}{\sum_{a_o\in A'}\left|s(a_i,a_o)\right|} \qquad (3\cdot 1)$$

ただし，A_i' は $a_o \in neighbor(a_i)$ の集合，$\bar{r}_i$ は全アイテムの $r_{i,k}$ の平均である．

最後に，予測評価値 p_i に基づいて上位 N 個の推薦リスト $L_{p_i}:\{1,2,\cdots,N\}\rightarrow B$ が計算される．関数 L_{p_i} は最も高い予測値を持つアイテムを 1 位とした降順の推薦ランキングを示す．

6-3-4 情報推薦の未来

嗜好学習と情報抽出の分野における，今後の研究課題はいくつか存在する．一つは，ユーザの嗜好や興味の変化や，短期的な嗜好・興味への対応である．これには，閲覧・購買行動の時系列的な情報を用いたり，より詳細にユーザの行動を記録したりするなどのアプローチが考えられる．もう一つは，推薦結果の評価に関するものである．従来の情報推薦は，情報検索の分野で用いられる精度・再現率を用いた評価を行ってきた．しかし，商用目的では，推薦の新規性や意外性なども考慮しなければならない．正確性を犠牲にしても，推薦が均一化しないようにする方法が必要となる．

■参考文献

1) T.W. Malone, et al. :“Semi-structured Messages are Surprisingly Useful for Computer-Supported Coodination,” Proc. of CSCW’86, pp.102-114, 1986.
2) P.W. Foltz :“Using Latent Semantic Indexing for Information Filtering,” Proc. of ACM Conference on Office Inforamtion Systems, pp.40-47, 1990.
3) P. Resnick, et al. :“GroupLens : An Open Architecture for Collaborative Filtering of Netnews,” Proc. of CSCW’94, pp.175-186, 1994.
4) S. Loeb and D. Terry (ed.) :“Information Filtering,” Commu. ACM, vol.35, no.12, pp.26-81, 1992.
5) P. Resnick and H. Varian (ed.) :“Recommender Systems,” Commun. ACM, vol.40, no.3, pp.56-89, 1997.

6) D. Riecken (ed.)：“Personalized Views of Personalization,”Commun ACM, vol.43, no.8, pp.26-158, 1992.
7) 土方嘉徳：“情報推薦・情報フィルタリングのためのユーザプロファイリング技術,”人工知能学会誌，vol.19, no.3, pp.365-372, 2004.

■7 群－6 編－6 章

6-4 時系列解析，パターン発見，変化点検出

（執筆者：井手　剛）［2009 年 1 月 受領］

本節では主に実数値の時系列データからの知識発見技術について解説する．時系列的に観測されるデータについて，過去データから未来を予測したいというのは極めて自然な問題設定であり，これまで膨大な研究の蓄積がある．一方，90 年代半ばから，分類やクラスタリングといった，もともとは独立同一分布（individually and identically distributed：i.i.d.）に従う標本を対象にした機械学習の手法が盛んに研究されるようになり，時系列データからの知識発見技術に興味が持たれるようになった．すなわち，時系列マイニングとは，時系列性と i.i.d. 性の交差するところに生まれた新しい研究領域である．本節では，伝統的な時系列解析手法に簡単に触れた後，時系列データからのパターン発見のための基本タスクを概観する．

6-4-1 時系列解析

時系列解析の基本的な問題設定は大きく 2 つある．一つは過去のデータを基に未来を予測する問題であり（予測），もう一つは，その予測モデルのパラメータを既知のデータから求めること（システム同定）である．例として，M 次の自己回帰モデル（Autoregressive model：AR model）

$$x^{(t)} = a_1 x^{(t-1)} + a_2 x^{(t-2)} + \cdots + a_M x^{(t-M)} + \varepsilon$$

を考える．ここで $x^{(t)} \in \mathbb{R}$ は時刻 t での観測値，ε は観測ノイズを表す．このモデルにおけるシステム同定問題とは，観測データから自己回帰係数 $\{a_s | s = 1, \cdots, M\}$ を求めることである．一般に，これは最尤推定の問題として定式化される [23),22)]．自己回帰係数が求まれば，時系列予測は，過去の M 個のデータ点を使って自明に実行できる．

隠れマルコフモデル（Hidden Markov Model：HMM）および状態空間モデル（state-space model）は，非観測の潜在状態を取り入れることで自己回帰モデルを一般化したものであり，強力なモデリング能力と広範な応用を持つ [8),16),20),21)]．両者の相違は，潜在状態が離散的（HMM）か，連続的か（状態空間モデル）という点にある．自己回帰モデルと異なり，これらのモデルでは予測問題は自明には解けない．潜在状態の推定が必要となるからである．その目的のために，HMM においてはビタビ（Viterbi）アルゴリズム，ガウスノイズのもとでの状態空間モデルにおいてはカルマン（Karman）フィルタという計算法が知られている [3)]．

データマイニングの文脈では，時系列の予測というよりも，モデルを同定した後に，そのモデルを次の処理に使うことも多い．いわゆるモデルベースの方法である．最初期の仕事の一つとして，スミス（Smyth）による HMM ベースの時系列クラスタリング [18)] を挙げておく．

6-4-2 パターン発見

本項では，時系列特有のタスクである，時系列整列（alignment），部分時系列クラスタリング，異常部位（discord）検出，頻出部位（motif）検出について概観する．

(1)　時系列整列

系列の整列（alignment）とは，複数の時系列の重なりが最大になるように重ね合わせることを指す．実数値時系列の場合は，重ね合わせの際に，時間軸をゴムのように伸縮させることを

許す．そのための手法は，動的時間伸縮法（Dynamic Time Warping：DTW）と呼ばれており，その数学的本質は動的計画法である．DTW はもともと音声認識の分野で提案され [17]，次いで 90 年代半ばにデータマイニングへの応用が提案された [2]．その後，様々な高速計算手法が提案され [11),12)]，現在では時系列データ同士の標準的な類似度計算手法として定着している．

(2) 部分時系列クラスタリング

ダス（Das）らの論文 [4] 以降，滑走窓（sliding window）で生成した部分系列に k 平均クラスタリングを適用することで，時系列に頻出する特徴的なパターンを見出す試みが数多く行われた（**図 4･1** 参照）．しかし今では，実験的 [13] および理論的に [5]，パターン発見手法としての，その手法の無効さが示されている．

図 4･1　幅 w の滑走窓による特徴ベクトル（部分系列）の生成

(3) 頻出部位検出，異常部位検出

部分時系列クラスタリングでは，クラスタ内の平均ベクトルを抽出パターンとみなした．これは滑走窓によるいわば「並進対称性の呪い」の悪影響を受けるため [5]，頻出部位（motif）検出というタスクが提案された [15]．滑走窓により部分系列を生成し，すべての部分系列同士の距離を計算する．あるパラメータ ε を与えたとき，第 1 頻出部位とは，ε-近傍の個数が最も多い部分系列のことを指す．ε-近傍数の順に，第 2 頻出部位なども定義される．

一方，異常部位（discord）検出というタスクでは，各部分系列の k-近傍を求め，その k 個の距離の中で最も大きいものをその部分系列のスコアとする．スコア上位のものが異常部位とみなされる．

異常部位および頻出部位検出タスクでは，部分系列の数を m としたとき，$O(m^2w)$ の計算量が必要である．現在のところ，時系列を離散近似する方法が，とりわけユークリッド距離の下限を与えるという点から，実用的に有効だと主張されている [14]．

6-4-3 変化点検出

時系列データの変化点とは，データの生成モデルが変化した時点として定義される．$x \in \mathbb{R}$ の生成モデルを $p(x|\theta)$ と表したとき，時刻 t においてモデルパラメータが $\theta = \theta_0$ から θ_1 に変わったとすれば，その時刻が変化点である．この定義から示唆されるように，変化点検出問題は，伝統的には尤度比検定の問題として定式化されてきた [1]．尤度比は，ネイマン-ピアソンの補題の意味で最適な統計量であること，その対数が漸近的にカイ 2 乗分布に従うこと，など顕著に優れた性質を持ち，変化点検出理論の確固たる基盤をなしている．

しかし実用上は，データの生成モデルの推定が簡単でないこと，多くの場合，漸近分布は役に立たないこと，などのため，文脈に応じて様々な工夫がされてきた．最近の代表的な仕事と

しては，逐次忘却型自己回帰モデルによるもの [19]，逐次的主成分分析によるもの [6),7)]，状態空間モデルの同定によるもの [10)]，密度比の直接推定によるもの [9)]，などが挙げられる．詳細はこれら論文および成書 [1)] を参照されたい．

■参考文献

1) M. Basseville and I.V. Nikiforov：Detection of Abrupt Changes. Prentice Hall, Englewood Cliffs, New Jersey, 1993.
2) D. Berndt and J. Clifford：“Using dynamic time warping to find patterns in time series,” AAAI-94 Workshop on Knowledge Discovery in Databases, AAAI, 1994.
3) Christopher M. Bishop：Pattern Recognition and Machine Learning, Springer-Verlag, 2006.
4) Gautam Das, King-Ip Lin, Heikki Mannila, Gopal Renganathan, and Padhraic Smyth：“Rule discovery from time series,” Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining, 1998.
5) Tsuyoshi Idé：“Why does subsequence time-series clustering produce sine waves?,” Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases, pp.311-322, 2006.
6) Tsuyoshi Idé and Keisuke Inoue：“Knowledge discovery from heterogeneous dynamic systems using change-point correlations,” In Proc. SIAM Intl. Conf. Data Mining, pp.571-575, 2005.
7) Tsuyoshi Idé and Koji Tsuda：“Change-point detection using krylov subspace learning,” Proc. 2007 SIAM Intl. Conf. Data Mining, pp.515-520, 2007.
8) Frederick Jelinek：Statistical Methods for Speech Recognition, MIT Press, 1998.
9) Yoshinobu Kawahara and Masashi Sugiyama：“Change-point detection in time-series data by direct density-ratio estimation,” Proc. SIAM Intl. Conf. Data Mining, 2009.
10) Yoshinobu Kawahara, Takehisa Yairi, and Kazuo Machida：“Change-point detection in time-series data based on subspace identification,” Proc. IEEE Intl. Conf. Data Mining, pp.559-564, 2007.
11) E. Keogh and M. Pazzani：“Scaling up dynamic time warping for data mining applications,” In Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining, pp.285-289, 2000.
12) Eamonn Keogh：“Exact indexing of dynamic time warping,” Proc. Intl. Conf. Very Large Data Bases, pp.406-417, 2002.
13) Eamonn Keogh, Jessica Lin, and Wagner Truppel：“Clustering of time series subsequences is meaningless: Implications for previous and future research,” Proc. IEEE Intl. Conf. on Data Mining, pp.115-122. IEEE, 2003.
14) Eamonn J. Keogh, Jessica Lin, and Ada Wai-Chee Fu：“HOT SAX: Efficiently finding the most unusual time series subsequence,” In Proc. IEEE Intl. Conf. Data Mining, pp.226-233, 2005.
15) Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Pranav Patel：“Finding motifs in time series,” Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining, pp.53-68, 2002.
16) Christopher D. Manning and Hinrich Schuetze：Foundations of Statistical Natural Language Processing, MIT Press, 1999.
17) H. Sakoe and S. Chiba：“Dynamic programming algorithm optimization for spoken word recognition,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol.26, no.1, pp.43-49, 1978.
18) Padhraic Smyth:“Clustering sequences with hidden markov models,”Advances in Neural Information Processing Systems 9, pp.648-655, Cambridge, MA, 1997. MIT Press.
19) K. Yamanishi and J. Takeuchi:“A unifying framework for detecting outliers and change points from nonstationary time series data,” Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining, pp.676-681, 2002.
20) 赤池弘次，北川源四郎（編）：時系列解析の実際 I，朝倉書店, 1994.
21) 赤池弘次，北川源四郎（編）：時系列解析の実際 II，朝倉書店, 1995.
22) 片山 徹：システム同定入門，朝倉書店, 1993.
23) 北川源四郎：時系列解析入門，岩波書店, 2005.

■7 群－6 編－6 章

6-5 外れ値検出・不正検出

（執筆者：山西健司）［2008 年 12 月 受領］

セキュリティの分野では，攻撃や侵入や情報漏えいなどの不正行為を早期に検出することが重要である．このような事象は，その手口が決まっていれば，アクセスログを調べて，それを署名（シグネチャ）と呼ばれるパターンを作り上げ，それとのパターンマッチングにより検出すればよい．そのような方法を署名ベースの検出と呼ぶ．しかしながら，この方法では，パターンの定まらない攻撃や，未知のタイプの攻撃が来たときには検出できないといった問題が起こる．そこで，過去の大量のログからパターンを学習し，これから大きく外れた異常を検出することにより，適応的に未知のタイプの攻撃を検出することができる．これを学習型の外れ値検出（outlier detection）と呼ぶ．これは，障害検出や故障検出の分野で，未知のタイプの障害や故障を検出する際にも有効である．

外れ値検出には，マハラノビス距離に基づく方法，Nearest Neighbor 法に基づく方法，クラスタリングに基づく方法，One-class SVM に基づく方法，確率密度推定に基づく方法などがある [3]．

6-5-1 マハラノビス距離に基づく外れ値検出

まず，最も単純な統計的な外れ値検出方法である，マハラノビス距離に基づく外れ値検出について簡単に説明しよう．仮に，データは n 次元連続値ベクトルであるとして，これまで得られたデータ列を $x^n = x_1, \cdots, x_m$ とし，i 番目のデータは $x_i = (x_{i,1}, \cdots, x_{i,n})$ と記すとき，その平均値ベクトルを μ，分散共分散行列を Σ は以下で求められる．

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i, \qquad \Sigma = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)^T(x_i - \mu)$$

そこで，θ を閾値パラメータとして，新しいデータ x に対して

$$\left\{(x-\mu)^T \sum^{1} (x-\mu)\right\}^{1/2} > \theta \qquad (5 \cdot 1)$$

を満たすならば x は外れ値であると判定する．式(5･1)の左辺は x と μ のマハラノビス距離と呼ばれるものである（例えば，参考文献 1)参照）．

6-5-2 確率密度推定に基づく学習型外れ値検出

上記の閾値法マハラノビス距離に基づく外れ値検出が有効に機能するためには，「データが定常的に同一のガウス分布 $N(\mu, \Sigma)$ から発生する」ことを前提としている．実際にはデータの発生分布は単峰のガウス分布であるとは限らず，しかも非定常性がある．このような場合に対応してデータの分布を適応的に学習し，リアルタイムに外れ値検出を実現するための方式 SmartSifter が提案されている [5]．

(1) SmartSifter の基本原理

SmartSifter は，基本的にはデータの統計的パターンを学習し，それに基づいて各々のデータをスコアリングする．これをオンラインで行う．その基本原理を以下にまとめる．

(a) データの発生機構を階層的な確率モデルで表現する

x は離散値変数ベクトル，**y** は連続値変数ベクトルとし，データを **(x, y)** のように表すとする．**(x, y)** の同時確率分布を $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})\, p(\mathbf{y} \mid \mathbf{x})$ のように記す．

ここで，$p(\mathbf{x})$ は有限個の排反なセルを持つヒストグラム型の確率密度関数を用いて表す．すなわち，**x** の空間を $\mathcal{X}$ とし，$\mathcal{X}$ を分割する排反なセル集合 $\mathcal{A}_1, \cdots, \mathcal{A}_M$ が与えられているとして，各セル上では一定の確率値をとるような $\mathcal{X}$ 上の確率分布 $p(\mathbf{x})$ を考える．

次に，$\mathcal{X}$ の各セルに対して，そこに入ったすべての **x** に対応する **y** の条件つき分布 $p(\mathbf{y}|\mathbf{x})$を，次式で与えられるガウス混合モデルを用いて表す．

$$p(\mathbf{y}|\mathbf{x} : \theta) = \sum_{i=1}^{k} c_i p(\mathbf{y}|\mu_i, \Lambda_i)$$

ここに，k はガウス混合分布の要素の数，$e_i \geq 0$，$\sum_{i=1}^{k} c_i$，各 $p(\mathbf{y}|\mu_i, \Lambda_i)$ は平均 μ_i，分散共分散行列 Λ_i の d 次元ガウス分布

$$p(\mathbf{y}|\mu_i, \Lambda_i) = \frac{1}{(2\pi)^{d/2}|\Lambda_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu_i)^T \Lambda_i^{-1}(\mathbf{y}-\mu_i)\right) \quad (i = 1, \cdots, k)$$

であるとする．d はデータの次元を表す．$\theta = (c_1, \mu_1, \Lambda_1, \cdots, c_k, \mu_k, \Lambda_k)$ とおく．

(b) 忘却型学習アルゴリズムで確率モデルを学習する

データ系列が $(\mathbf{x}_1, \mathbf{y}_1)$，$(\mathbf{x}_2, \mathbf{y}_2)$，… のようにオンラインで与えられる場合，$t$ 番目の入力データ $(\mathbf{x}_t, \mathbf{y}_t)$ が与えられたときに，まず $\mathbf{x}_t$ が入るセルを同定し，後述する SDLE（Sequentially Discounting Laplace Estimation）アルゴリズムを用いて **x** の分布を推定し，推定分布を $p^{(t)}(\mathbf{x})$と書く．

次に，そのセルについて **y** の分布であるガウス混合分布を後述する SDEM（Sequentially Discounting Expectation and Maximizing）アルゴリズムを用いて推定し，推定分布を $p^{(t)}(\mathbf{y} \mid \mathbf{x})$ と書く．他のセルについては，$p^{(t)}(\mathbf{y}|\mathbf{x}) = p^{(t-1)}(\mathbf{y}|\mathbf{x})$ とおく．

Given: **A partitioning of the domain** $\{\mathcal{A}_1, \cdots, \mathcal{A}_M\}$, r, **and** β.
初期化:
Let $T_0(i) = 0 \quad (i = 1, \cdots, M)$.
$t := 1$
パラメータ更新:
while $t \leq T$ （T: **サンプル数**)
Read $\mathbf{x}_t = (x_1, \cdots, x_n)$
For the i**-th cell**,

$$T_t(i) := (1-r)T_{t-1}(i) + \delta_t(i) \quad \text{(各セルの統計量の計数)}$$

$$q^{(t)}(i) := \frac{T_t(i) + \beta}{(1-(1-r)^t)/r + M\beta} \quad \text{(Laplace推定)}$$

For each $\mathbf{x} \in \mathcal{A}_i$,

$$p^{(t)}(\mathbf{x}) := q^{(t)}(i) \mid \mathcal{A}_i \mid \quad \text{(シンボルごとの確率推定)}$$

ここで，t 番目のデータが i 番目のセルに入ったら $\delta_i(i) = 1$ とし，そうでない場合は$\delta_i(i) = 0$ とおく．
$t := t + 1$

図 5・1　SDLE アルゴリズム

Given: r, α, k

初期化:

Set $\mu_i^{(0)}$, $c_i^{(0)}$, $\bar{\mu}_i^{(0)}$, $\Lambda_i^{(0)}$, $\bar{\Lambda}_i^{(0)}$ $(i = 1, \cdots, k)$.

$t := 1$

パラメータ更新:

while $t \le T$ (T: サンプル数)

Read $\mathbf{y}_t$

for $i = 1, 2, \cdots, k$

$$\gamma_i^{(t)} := (1-\alpha r)\frac{c_i^{(t-1)} p\left(\mathbf{y}_t \middle| \mu_i^{(t-1)},\ \Lambda_i^{(t-1)}\right)}{\sum_{i=1}^{k} c_i^{(t-1)} p\left(\mathbf{y}_t \middle| \mu_i^{(t-1)},\ \Lambda_i^{(t-1)}\right)} + \frac{\alpha r}{k}$$

$$c_i^{(t)} := (1-r)c_i^{(t-1)} + r\gamma_i^{(t)}$$

$$\bar{\mu}_i^{(t)} := (1-r)\bar{\mu}_i^{(t-1)} + r\gamma_i^{(t)} \cdot \mathbf{y}_i$$

$$\mu_i^{(t)} := \bar{\mu}_i^{(t)} / c_i^{(t)}$$

$$\bar{\Lambda}_i^{(t)} := (1-r)\bar{\Lambda}_i^{(t-1)} + r\gamma_i^{(t)} \cdot \mathbf{y}_t \mathbf{y}_t^T$$

$$\Lambda_i^{(t)} := \bar{\Lambda}_i^{(t)} / c_i^{(t)} - \mu_i^{(t)} \mu_i^{(t)^T}$$

$t := t + 1$

図 5・2　SDEM アルゴリズム

SDLE アルゴリズムは離散分布に対するラプラス推定のアルゴリズムをオンライン処理で，かつ忘却機能を持たせるように改良したものである．T は全データ数，$T(i)$ は i 番目のセルに入ったデータ数，β は正の定数として SDLE アルゴリズムはこのような確率分布を**図 5･1** のように学習する．$0 \le \mathrm{r} \le 1$ は忘却係数であり，r の値が小さいほど，過去を忘却する．

SDEM アルゴリズム θ を推定するのに，EM アルゴリズムをデータ入力ごとに過去のデータによる効果を徐々に減らしながら，オンラインで学習する．これを**図 5･2** にまとめる．ここで，$0 \le \mathrm{r} \le 1$ は忘却係数であり，r の値が小さいほど，過去を忘却する．$0 < \alpha < 1$ は与えられた定数である．各反復に対する SDEM アルゴリズムの計算時間は $O(d^3k)$ である．ここに d はデータの次元であり，k はガウス混合分布の要素の数である．

(c) 学習されたモデルを基にスコアを計算する

SmartSifter は各データに対して上で学習されたモデルに基づいてスコアを Hellinger スコアまたは対数損失で計算する．ここで $p^{(t)}(\mathbf{x}, \mathbf{y})$ を t 番目のデータから学習された確率分布とするとき，t 番目のデータに対する Hellinger スコアを以下で定める．

$$S_H(\mathbf{x}_t, \mathbf{y}_t) = \frac{1}{r^2} \sum_{\mathbf{x}} \int \left(\sqrt{p^{(t)}(\mathbf{x}, \mathbf{y})} - \sqrt{p^{(t-1)}(\mathbf{x}, \mathbf{y})} \right)^2 d\mathbf{y}$$

また，対数損失（Shannon 情報とも呼ぶ）を以下で定める．

$$S_L(\mathbf{x}_t, \mathbf{y}_t) = -\ln p^{(t-1)}(\mathbf{x}_t, \mathbf{x}_t)$$

Hellinger スコアは分布 $p^{(t)}$ が $(\mathbf{x}_t, \mathbf{y}_t)$ からの学習によって，$p^{(t-1)}$ からどれくらい大きく動いた

かを学習前後の確率分布の Hellinger 距離で示すものである．したがって，高いスコアのデータは確率モデルの変化に大きく寄与したという意味で，外れ値であるみなすことができる．一方，対数損失はデータの過去のモデルに対してデータの意外性としての意味を持つ．

SmartSifter の使い方としては，与えられたデータセットに対して，各データに逐次的にスコアを与え，すべてのデータのスコアリングが終わった後，スコアに従ってデータをソートして，外れ値上位リストを出力するといった使い方が一般的である．

(2) SmartSifter の応用

SmartSifter のような適応的外れ値検出は侵入検出に応用されている [3),5)]．実際，KDDCup'99 と呼ばれる通信ログのデータ集合から外れ値を検出することにより，ランダム検査に較べて数十倍～100 倍高い精度で未知の侵入を検出できることが示されている [5)]．更に，SmartSifter はシステムの性能データからの障害検出にも応用されている [2)]．また，大量のレセプトデータからの外れ値検出により，不審な医療機関の同定などを行っている例もある [5)]．

6-5-3 異常行動検出

上述の適応的外れ値検出では，一つひとつが多次元データであるとして，そのパターンを学習し，外れ値を検出した．一方で，一つひとつのデータの単位が時系列（セッションと呼ぶ）として，異常なセッションを検出する技術を異常行動検出と呼ぶ．

異常行動検出の方法の一つとして，セッションの生成モデルを隠れマルコフモデルの混合モデルとして表し，この学習に基づいて異常行動検出を行う方法がある [4)]．そこでは，独立な複数のセッションを入力とし，SmartSifter と同様にオンライン忘却型の学習アルゴリズムを用いてセッションのパターンを学習し，Hellinger スコアや対数損失を用いてセッションの異常をスコアリングする．

本技術は，UNIX のコマンド履歴からのなりすまし者の検出や，システムの SYSLOG と呼ばれるシステム動作履歴からの障害検出などに適用され，障害の予兆を検出するのに有効であることが示されている [4)]．

■参考文献

1) V. Barnett and T. Lewis：Outliers in Statistical Data, John and Wiley & Sons, 1994.
2) H. Chen, G. Jiang, K. Yoshihira：“Monitoring high-dimensional data for failure detection and localization in large-scale computing systems,”IEEE Trans. on Knowledge and Data Engineering, vol.20, no.1, pp.13-25, Jan., 2008.
3) A. Lazarevic, L. Ertoz, A. Ozgur, J. Srivastava, and V. Kumar：“A comparative study of anomaly detection schemes in network intrusion detection,”Proceedings of The Third SIAM Conference on Data Mining (SDM03), 2003.
4) K. Yamanishi and Y. Maruyama：“Dynamic syslog mining for network failure monitoring,”Proceedings of The Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD05), ACM Press, pp.499-508, 2005.
5) K. Yamanishi, J. Takeuchi, G. Williamas, and P. Milne：“On-line unsupervised oultlier detection using finite mixtures with discounting learning algorithms,”Data Mining and Knowledge Discovery Journal, pp.275-300, vol.8, Issue 3, 2004.