

混合多項分布推定を用いた肝炎データにおける 異常検査値の類型化

渡辺 健志

nabekun@slab.dnj.ynu.ac.jp

鈴木 英之進

suzuki@dnj.ynu.ac.jp

横浜国立大学大学院工学府物理情報工学専攻電気電子ネットワークコース

概要

本論文は医療検査データをトランザクションデータと見なし、EM法による混合多項分布推定を用いて異常検査値に関する基本的な類形を導出する。トランザクションデータは、アイテムを属性とする表形式データと見なした場合疎なデータに相当し、解析が困難である場合が多い。EM法による混合多項分布推定は、多項分布で表される基本的な類形を高速に求めるため、トランザクションデータを属性値の分布が偏っていない表形式データに変換する基盤手法として有望である。肝炎データを用いた実験の結果、求められた類形の中には医学的な意味が明確なものも存在し、この手法により検査結果の傾向に関する可読性が向上する事が確認された。

Prototyping Abnormal Medical Test Values in Hepatitis Data with Mixture Multinomial Distribution Estimate

Takeshi Watanabe Einoshin Suzuki

nabekun@slab.dnj.ynu.ac.jp suzuki@dnj.ynu.ac.jp

Department of Electrical and Computer Engineering, Graduate School of Engineering,
Yokohama National University.

Abstract

This paper regards medical test data as transactional data, and induces basic prototypes of abnormal medical test values using mixture multinomial distribution estimate by the EM method. Viewed as a table-formatted data set with items as attributes, a transactional data set corresponds to a sparse data set, and is typically hard to be analyzed. Mixture multinomial distribution estimate by the EM method can be considered as promising as foundation of a method to transform such a data set into a table-formatted data set of which value distribution is not skewed since the method rapidly obtains prototypes each of which is represented by a multinomial distribution. Experimental results with hepatitis data show that some of the obtained prototypes have clear meaning in medicine, and this method improves readability of tendencies on medical tests.

1 はじめに

近年の高度情報化社会において、ハードウェアの低価格化や種々の情報の電子化にともないデータベースはますます大規模なものとなってきた。しかし膨大な量のデータの解析は人の処理能力をはるかに越えるものとなってしまう、現在データベースは有効に活用されているとはいいがたい。そのため、計算機による有効な知識発見が必要とされている。大量のデータは個々に扱うより、類似したものをまとめて扱う方がデータを大域的に調べられる。そのためプロファイリング [1] など、データから傾向パターンを抽出して類型を作成する多くの研究が行われている。

マーケットバスケットデータに代表されるトランザクションデータは、取引における 1 回の処理を 1 トランザクションとして記録したデータであり、多数のアイテムから構成される。しかし多くの場合において 1 トランザクションには少数のアイテムしか現れない。トランザクションデータをアイテムを属性とする表形式データとして表した場合、このデータの値はほとんどが 0 でごくわずかが 1 と疎になり可読性が低く、解析が困難である。そのためトランザクションデータを類型化し、各トランザクションをその類型に基づいてクラスタリングすることで密なデータとすることが考えられる。この結果、可読性が向上され、解析が容易になると考えられる。

医療診断における血液検査などの多数の項目を調べる検査では、各項目ごとに正常値範囲が定められており、この範囲に入らない検査値は異常値とされ、患者の状態を診断する際に重要な要素となる。検査項目をアイテム、異常な検査値をアイテム購入と見なすと、検査データは内容が疎な離散的データとなり、トランザクションデータと見なせる。ここで多項分布 [2] は、独立な試行によって起きる複数の排反な事象に対する確率分布であり、プロファイリング [1] に用いられているように、トランザクションデータに適している。よって本研究では、検査データをトランザクションデータと見なし、混合多項分布を用いて表現した異常値を持つ検査項目の類型を導出する。実験では実データとして肝炎データを用い、EM アルゴリズム [3]

によって基本的な類型を導出した。

2 混合多項分布

トランザクション数 M 、アイテム数 c のトランザクションデータ $T = (x_1, x_2, \dots, x_M)$ が与えられたとする。個々のトランザクション x_i は、それぞれ項目を表すアイテム $a_1, a_2, \dots, a_c$ から構成され $x_i = (n(a_{i1}), n(a_{i2}), \dots, n(a_{ic}))$ と表される。ただし a_{ij} は i 番目のトランザクションの j 番目のアイテム値である。

多項分布は、例えばサイコロを複数回投げたとき各面が何回ずつ出るかというように n 回の試行が独立で、各回の試行によって m 個の排反な事象のうちのどれかが起こる場合の確率分布である。混合多項分布とは多項分布で表される複数の基本パターンを持ち、各基本パターンが生起確率に従って起きる確率分布モデルである。例えるならば、多項分布とは各面がそれぞれ生起確率を持つ多面体のサイコロを複数回投げた場合の各面が出る回数に関する確率を表す。一方、混合多項分布は、そのサイコロが複数種類用意され、使用するサイコロの選択も確率によって決められる事象を表す。

アイテム $a_1, a_2, \dots, a_c$ が、それぞれ $n(a_1), n(a_2), \dots, n(a_c)$ 回起こる確率 $P(n(a_1), n(a_2), \dots, n(a_c) | p_1, p_2, \dots, p_c)$ は、各事象の起きる確率を $p_1, p_2, \dots, p_c$ とすると、

$$P(n(a_1), n(a_2), \dots, n(a_c) | p_1, p_2, \dots, p_c) = \frac{N!}{n(a_1)!n(a_2)! \dots n(a_c)!} p_1^{n(a_1)} p_2^{n(a_2)} \dots p_c^{n(a_c)} \quad (1)$$

で与えられる。ただし N は全試行回数である。

また混合モデルは以下で定義される。

$$p(x) = \sum_{k=1}^K p(x|k)p(k) \quad (2)$$

ただし x はデータベクトルであり $x = (n(a_1), n(a_2), \dots, n(a_c))$ と表される。 k はある基本パターンであり $k = (p_{k1}, p_{k2}, \dots, p_{kc})$ と表される。ただし p_{ij} は基本パターン i の j 番目のアイテムの生起確率である。 $p(k)$ は基本パターン k の生起確率、 $p(x|k)$ は基本パターン k のときに x となる条件つ

き確率であり,

$$p(x|k) = P(n(a_1), n(a_2), \dots, n(a_c) | p_{k1}, p_{k2}, \dots, p_{kc}) \quad (3)$$

と表される.

3 EM アルゴリズム

EM アルゴリズムでは山登り法により, 混合モデルの最尤パラメータを算出する. 2章で定義したトランザクションデータ T が与えられた場合, K 個の基本パターンからなる混合モデルの負の対数尤度は

$$\varepsilon = - \sum_{m=1}^M \ln \left(\sum_{k=1}^K p(x_m|k)p(k) \right) \quad (4)$$

で求められ, これを最小とする混合モデルを算出する. 手順は以下ようになる.

- 1) 各基本パターンのパラメータ初期値を決定する.
- 2) ベイズ則から事後確率 $p(k|x)$ を求める.

$$p(k|x) = \frac{p(x|k)p(k)}{p(x)} \quad (5)$$

本研究では基本パターンは多項分布で表すので, $p(x)$, $p(x|k)$ はそれぞれ式 (2), (3) で与えられる.

- 3) パラメータ更新

基本パターンの生起確率は

$$p^{new}(k) = \frac{1}{M} \sum_{m=1}^M p^{old}(k|x_m) \quad (6)$$

基本パターンの各アイテム生起確率は

$$p^{new}(a_i|k) = \frac{p(k|x_m)n(a_i)}{p(k|x_m) \sum_{j=1}^c n(a_j)} \quad (7)$$

で更新される.

以下, 収束するまで 2), 3) を繰り返す.

4 実験

4.1 条件

実データとして千葉大学病院第一内科第二研究室から提供していただいた B 型, C 型肝炎患者

データを用いる. このデータは検体検査結果情報, 肝生検情報, およびインターフェロン投与情報から構成される. 検体検査結果情報は肝炎患者の受けた血液検査や尿検査の日付と結果から構成され, 検査結果が高過ぎる場合は “H”, 低過ぎる場合は “L” と結果数値の後に記されている. 肝生検情報は肝生検¹ の日付や結果から構成され, 組織の繊維化状態と活動性は軽い順にそれぞれ F0~F4, A1~A3 で示されている. またインターフェロン投与情報にはインターフェロン² を投与した日時や回数記されている.

実験では検体検査結果情報を用いる. 患者 1 人の 1 回の検査を 1 トランザクション, 検査項目をアイテムと考える. つまりある患者が 1 日目に 2 回検査を受け, 2 日目に 3 回の検査を受けた場合, それぞれの検査は $t_1, t_2, \dots, t_5$ となる. 結果に異常 (H か L) がある場合その項目を 1 とし, それ以外を 0 とする. このデータはトランザクション数 58,716, アイテム数 458 から構成される. すなわち, 各基本パターンは肝炎において異常の起きる検査項目の傾向パターンを表すことになる.

EM アルゴリズムは初期値に近い局所解に収束する傾向があるので, 得た解は大域的な最適解とは限らない. そこで, 初期値をランダムに与えた試行を 100 回繰り返し, その中で最も尤度の高い混合モデルを採用する. 作成する基本パターン数は 2, 3, ..., 10 で行ったが, どのモデルを最適とするか判断が困難なので, 最も多い 10 パターンを記載する. なお収束精度は 0.001% とし, この精度に至らなくてもループ数が 100 になった場合は探索を終了した. 計算には CPU PentiumIII 1.26GHz のマシンを使用した.

求められた基本パターンは, 互いの類似度に基づきいくつかのグループに分かれると考えられる. 各基本パターンの類似度を測るため, 基本パターン同士の距離をダイバージェンス [4] によって計算する. ダイバージェンスは 2 つの確率分布の距離を測る関数で, 基本パターン k, l に対しては

$$D(k||l) = \sum_{i=1}^c p_{li} \ln \frac{p_{li}}{p_{ki}} \quad (8)$$

で定義される. ただし l は要素数であり, k, l は

¹ 肝臓の組織を採取し, 顕微鏡で調べる検査

² ウィルス性肝炎の特効薬的な薬

表 1: 10 パターンにおける結果. ただしトランザクション数は, その基本パターンに所属する確率が最も高いトランザクションの数

パターン	生起確率	トランザクション数 (割合)
1	17.65%	7457(12.70%)
2	22.50%	9788(16.67%)
3	4.70%	1342(2.29%)
4	11.44%	5552(9.46%)
5	11.08%	4728(8.05%)
6	5.33%	2087(3.55%)
7	0.51%	233(0.40%)
8	0.41%	138(0.24%)
9	1.55%	793(1.35%)
10	24.82%	26598(45.30%)

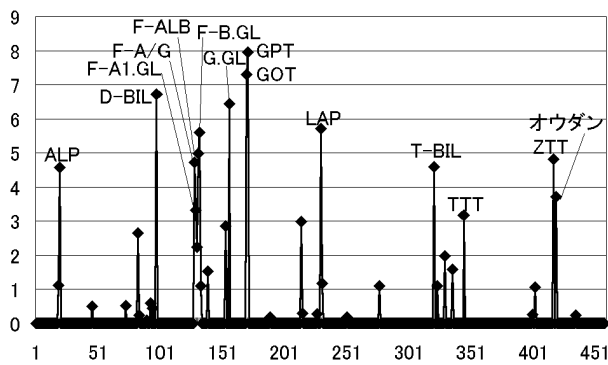


図 1: 基本パターン 1 の項目分布. ただし縦軸, 横軸はそれぞれ検査項目 i の生起確率 p_{1i} , 検査項目 i を表す

$k = (p_{k1}, p_{k2}, \dots, p_{kc})$, $l = (p_{l1}, p_{l2}, \dots, p_{lc})$ で表される確率分布である. 基本パターンに生起確率が 0 となるアイテムが現れた場合は計算不能に陥るので, その場合は生起確率 1×10^{-100} として計算する.

4.2 結果

作成されたモデルの基本パターン生起確率を表 1 に, 各基本パターン内のアイテム生起確率をグラフに表し図 1-10 に示す. 計算時間は約 11 時間であった.

基本パターンのグループ化にあたり, 各基本パターンの距離行列は非対称となるため, その平均距離が 10 以内であれば類似していると解釈した. その結果基本パターンは $\{1,2,3\}, \{4,5,6\}, 7,8,9,10$ の 6 グループに分かれた.

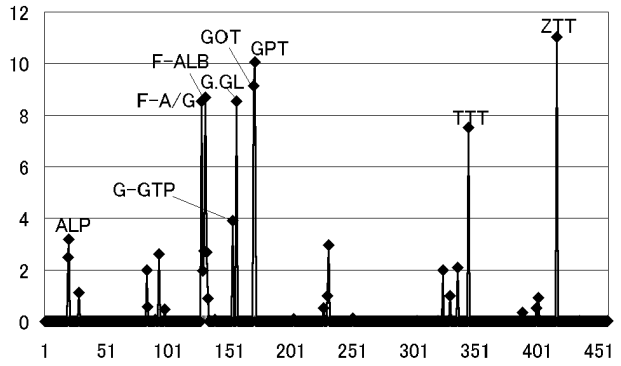


図 2: 基本パターン 2 の項目分布

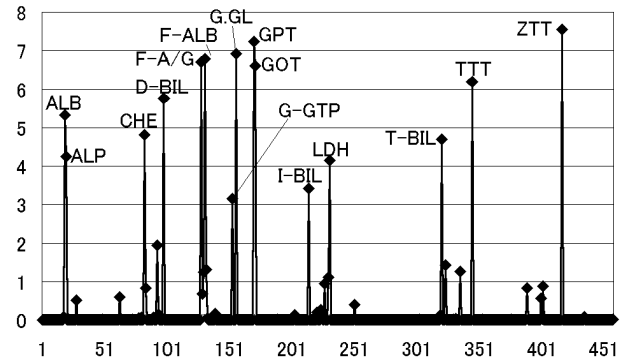


図 3: 基本パターン 3 の項目分布

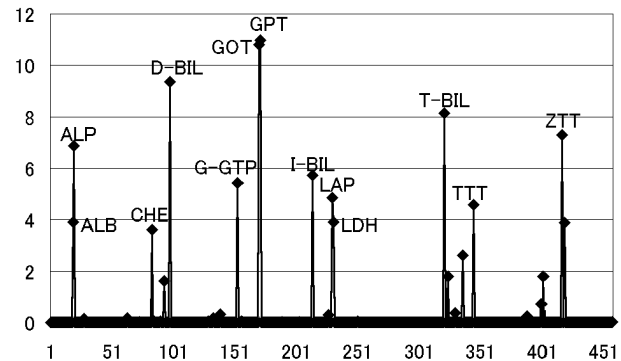


図 4: 基本パターン 4 の項目分布

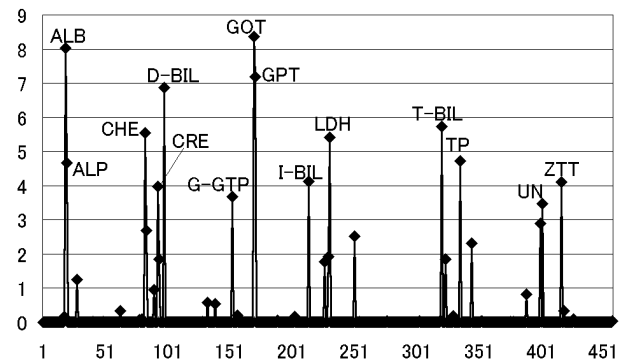


図 5: 基本パターン 5 の項目分布

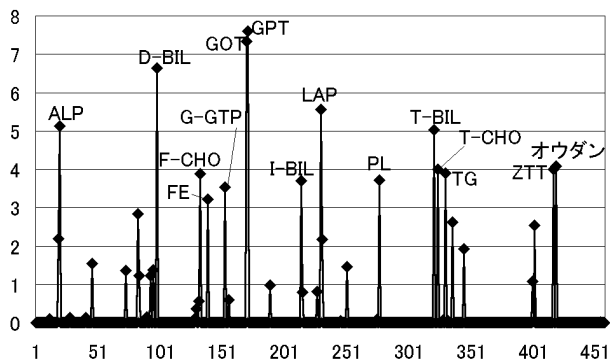


図 6: 基本パターン 6 の項目分布

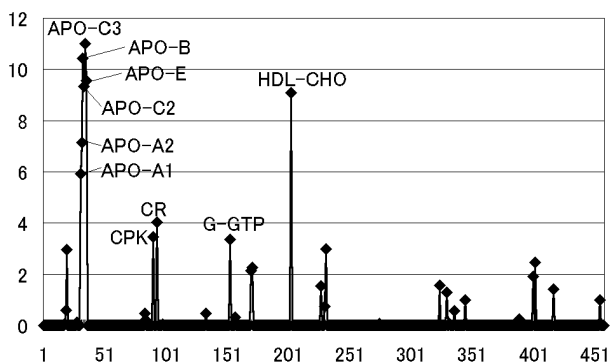


図 7: 基本パターン 7 の項目分布

4.3 考察

導出された基本パターンの内、他と大きく異なるものは7と9である。7はAPO³ 関連の項目が高い確率を示し、9は2項目が約70%と20%と非常に高い生起確率を持つ。専門家に意見をうかがったところ、基本パターン7は脂肪系蛋白に異常が起きている場合をよく表している、というコメントを頂いた。基本パターン9に所属するトランザクションは、1日に2回検査した場合における2回目の検査にほぼ占められる。これは再検査では検査する項目がケッチンやHBA1C/Xなどにほぼ決まっており、それらが基本パターンとしてとらえられたと考えられる。また基本パターン10には全トランザクションの半数近くが所属し、そのうち9578 トランザクションは異常となった項目がない。つまり基本パターン10は正常な結果に近いパターンを表していると考えられる。

肝生検における繊維化と活動性の検査、および

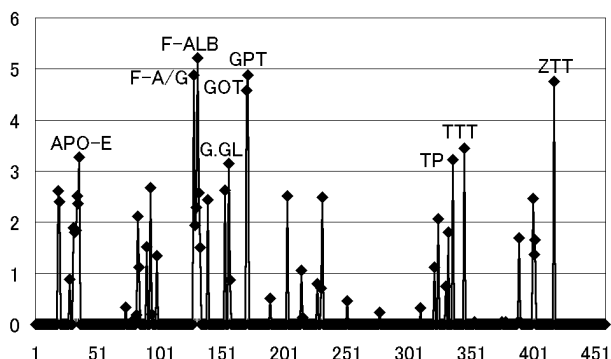


図 8: 基本パターン 8 の項目分布

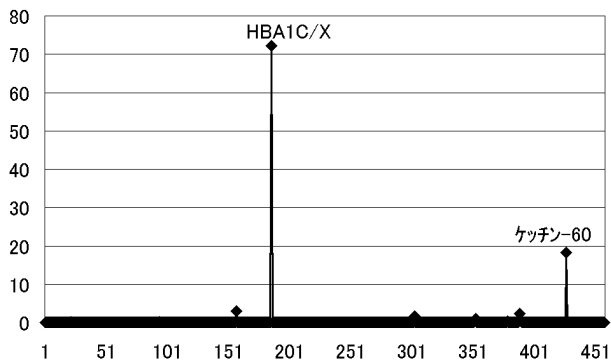


図 9: 基本パターン 9 の項目分布

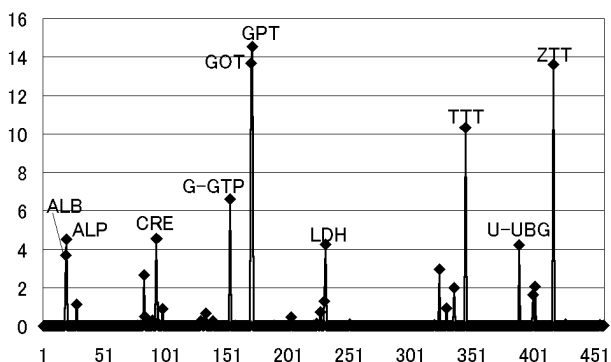


図 10: 基本パターン 10 の項目分布

³ アポ蛋白

インターフェロン投与を全て受けている 28 人中 10 人について、患者ごとに時系列を追って調べたところ、いくつか特徴的な検査結果が見つかったので、紹介する。この調査にあたり、肝生検情報とインターフェロン投与情報を併せて参考した。

1. 患者 ID: 87

インターフェロンを投与する前後 1 年間で行った検査は、全て基本パターン 10 に属すが、インターフェロンを投与している 5 ヶ月の間の検査では基本パターン 5 が混じり、異常検査項目として前後 1 年間では瀕出していた GOT, GPT⁴ があまり出現しなくなっている。インターフェロンの影響を大きく受けていると考えられる。

2. 患者 ID: 493

インターフェロンを投与している 6 ヶ月間で行った検査は、ほぼ基本パターン 10 に属すが、唯一 3 ヶ月目の検査で基本パターン 1, 2 に約 30% ずつ属す検査結果が見られた。この検査以外では常に正常検査項目であった U-UBG が異常となっている。インターフェロンの副作用が現れたと思われる。

3. 患者 ID: 702

肝生検までに行った検査は主に基本パターン 10 に属すが、肝生検翌日の検査は基本パターン 1, 4, 10 にそれぞれ約 20% ずつ属し、23 日後に行われた次の検査は基本パターン 4 に属す。肝生検までの検査では全ての検査において GOT, GPT は異常検査項目であったが、肝生検翌日の検査では GOT, GPT は正常検査項目となっており、次の検査では GPT だけが異常検査項目となっていた。肝生検によって何らかの影響を受けたものと思われる。

5 おわりに

今後の方針として、まず専門家に意見をうかがって手法に反映する。そして今回は異常検査値 “H”, “L” を持つ検査だけで実験したが、行った検査項目が異なる場合、全く違うパターンとなってしまう。すなわち検査した項目自体の影響を大きく受けてし

まう。そのため、行う検査項目群のパターンと各検査項目群内での異常検査項目を別々に扱う、2 重のクラスタリングを考えている。また、“H” と “L” は反対の意味を持つ場合もあるので区別し、検査値の数値の大小も考慮する。最終的には、異常検査基本パターンの時系列推移から肝生検の状態を予測するモデルを目標とする。

参考文献

- [1] I. V. Cadez, P. Smyth and H. Manilla: “Probabilistic Modeling of Transaction Data with Applications to Profiling, Visualization, and Prediction”, *Proc. Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 37–46, 2001.
- [2] 坂元 慶行, 石黒 真木夫, 北川 源四郎, 情報量統計学, 共立出版株式会社, pp. 12–15, 1993.
- [3] A.P. Dempster, N.M. Laird and D.B. Rubin, “Maximum likelihood from incomplete data via the EM Algorithm”, *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.
- [4] 有本卓: 確率・情報・エントロピー, 森北出版, pp. 33–39, 1980.

⁴ 肝臓の障害に敏感な酵素