D-025

Eigen Co-occurren Matrix(ECM) 手法を用いた医療データにおける患者のリスク評価と特徴抽出

Risk Evaluation and Feature Extraction of Patients in Medical Database Using Eigen Co-occurrence Matrix Method

小磯 知之[†]　岡 瑞起[†]　加藤 和彦[‡]

Tomoyuki Koiso[†]　Mizuki Oka[†]　Kazuhiko Kato[‡]

1. はじめに

医療分野において、様々な患者のデータ間の関係を見出し、病気の原因を探ることは、病気の治療・予防法を見つけるための研究として、盛んに行われている。現在殆どのこの分野の研究は、ロジスティック回帰分析を用いて行なわれている [2]。ロジスティック回帰分析は、ロジット関数と呼ばれるある疾病が起きる確率と起きない確率の比を表す関数を、説明変数 $\mathbf{X}=\{x_1, x_2, \cdots, x_n\}$ の 1 次式でモデル化したものである。この手法では、説明変数の項の和という形でデータ間の関係をモデル化している。この手法の問題点は、データ間の関係を求めたり、あるデータ間の関係が大きな影響を及ぼしている等の知見を得たりすることができないことである。

そこで我々はロジスティック回帰分析での問題点を解決するため、Eigen Co-occurrence Matrix(ECM) 手法 [3][4] を提案する。ECM 手法は、データ間の特徴的な関係を抽出し、それらをネットワーク図 (グラフ) として可視化することができる。

2. 提案手法

患者から得られるデータの間に関係を見出す為には、まず、データ間の関係をモデル化し、次に得られるデータの特徴抽出をして、最後にその特徴を医師が分析し易い形で可視化する必要がある。以下に我々の提案する ECM 手法のモデル化、特徴抽出、可視化の手順を示す。

2.1 モデル化

項目別に計測されたデータの項目間の関係性をモデル化する。まず、患者のデータの項目毎に健康度を数値化する。例えば健康である状態を-1 で表し、非健康である状態を 1 で表すこととする。健康度に変換されたある項目 $Factor_i$ と $Factor_j$ との関係を以下の様にモデル化する。

$$Factor_i + \alpha \times Factor_j \tag{1}$$

ここで、α は $Factor_i$ に対する $Factor_j$ の関係の強さを表す係数であるとする。この係数は任意に決定でき、例では簡単のために $\alpha = 1$ とした。患者 A のデータとその数値化の例をそれぞれ表 1、2 に示す。M 個の項目全てに対してこの操作を行うことで、ある患者のデータ間の関係を表した共起行列 (Co-occurrence Matrix) を生成する。患者 A のデータから生成された共起行列を表 3 に示す。

項目	酒	煙草	BMI	運動
患者 A	1000ml/日	0/日	30.0	無し

表 1: 患者のデータ

項目	酒	煙草	BMI	運動
患者 A	1	-1	1	1

表 2: 健康度の割り当て

2.2 特徴抽出

患者のデータから得られた $M \times M(\equiv L)$ 次元の共起行列に対して主成分分析を行って特徴を抽出する。主成分分析とは多変量で表されるデータの統計から、一次結合で表される新たな変量を構成し、互いに無相関な「主成分」に要約する手法である。このとき主成分を表す軸である固有ベクトルの集合 $\mathbf{a} = \{a_1, a_2, \cdots, a_L\}$ の各固有ベクトルから生成した $M \times M$ 次元行列を固有共起行列 (Eigen Co-occurrence Matrix) と呼ぶ。

ある共起行列 $\mathbf{x}$ に対する特徴ベクトル $\mathbf{C} = \{c_1, c_2, \cdots, c_L\}$ を $\mathbf{x}$ と $\mathbf{a}$ の内積を計算することにより求める。$\mathbf{C}$ の各成分は、共起行列 $\mathbf{x}$ を表す為の各固有共起行列の貢献度を表すことになる。

2.3 可視化

元の共起行列 $\mathbf{x}$ の近似行列 $\tilde{x}$ を、固有共起行列と特徴ベクトルから再構成する。この近似行列からネットワーク図を作ることで可視化を行う。固有共起行列を構成するための固有ベクトルを選ぶ個数 N を小さくするにより、固有共起行列 $\mathbf{a}$ とそれに対応する特徴ベクトル $\mathbf{C}$ を用いて元の共起行列を

$$\tilde{x} = \sum_{i=1}^{N} c_i a_i \quad = \sum_{i=1}^{N} \{X_i + Y_i\} = \sum_{i=1}^{N} X_i + \sum_{i=1}^{N} Y_i \tag{2}$$

のように低次元で近似して表現することができる。N を大きくするとより詳細な情報を得ることが出来る。こう

	酒	煙草	BMI	運動
酒	2	0	2	2
煙草	0	-2	0	0
BMI	2	0	2	2
運動	2	0	2	2

表 3: 共起行列

[†]筑波大学大学院
[‡]筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻/JST CREST

して近似された共起行列によって表されるネットワークは、元の共起行列の部分ネットワークではなく、固有共起行列から発生する全体的な構造を持つ近似ネットワークである。

さらに、近似した共起行列の非健康的な要素を表す正値 X_i と健康的な要素を表す負値 Y_i とに分けてネットワークを作ることによって、特徴を分離した形で可視化する。可視化された特徴的な関係から危険因子を推定することで、患者の病気に対するリスク評価や治療に役立てることができる。

3. 提案手法の適用

上述の提案手法を国際会議 ECML/PKDD Discovery Challenge 2004[1] で公開されている STULONG データセットに対して適用し、心疾患における特徴的な項目間の関係性を抽出し可視化する実験を行った。

3.1 データセット

STULONG データセットには Entry, Control, Letter, Death の 4 つのデータセットがあり、今回は Entry データセット中の RiskGroup のデータを用いた。RiskGroup とは [1] で定義されている心疾患の危険因子を少なくとも一つ持っている患者のグループである。Risk Group は Healthy Group(心疾患を患わなかったもの), Disease Group(心疾患を患った者), Dead Group(死亡した者) の 3 つから構成されている。この実験の目的は Healthy Group(2 4 2 人) と Disease Group(4 2 3 人) それぞれの項目間の関係からグループ間の差異を見出し、有用な考察を得ることである。

3.2 手順

実験では Healthy Group と Disease Group をそれぞれ学習用データと評価用データに 2 等分した。また、実験結果の有用性を、特徴ベクトル間の距離に基づいて評価用データの識別を行い、その識別率を計算することで行った。実験手順を以下に示す。

1. 特徴ベクトルの抽出

   各患者の学習用データから固有共起行列を生成し、特徴ベクトルを抽出した。同様に、評価用データからも特徴ベクトルを計算した。

2. 特徴抽出の検証

   抽出した特徴ベクトルが効果的に特徴を捉えられているかを、評価用データの特徴ベクトルと学習用データの特徴ベクトルとのユークリッド距離を計算し、評価用データを Healthy Group(○), Disease Group(×) でプロットした。結果を図 1 に示す。

3. ネットワーク図の構成

   特徴ベクトルから近似した共起行列を再構成し、得られた共起行列から絶対値の大きい順に値を取り出した。さらに、それらから Healthy, Disease の各グループの正 (健康的)、負 (非健康的) それぞれの値ごとに項目間の関係を示す 4 つのネットワーク図を作成した。例として Healthy Group の健康的な関係を取り出したネットワーク図を図 2 に示す。

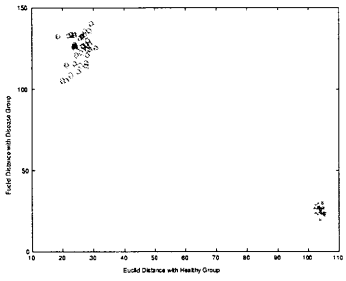

図 1: 評価用データのプロット

図 2: Healthy Group の負 (健康的) の関係

Healthy Group の健康な関係を示すネットワークは、健康を保つ事に大きく影響している項目間の関係を示している。また非健康な関係を示すネットワークは病気の原因ではない項目間の関係を示す。つまり、非健康な項目間の関係を持ちながらこのグループの患者は健康を保っているのである。

Disease Group の健康な関係を示すネットワークは病気に関与していない項目間の関係を示す。反対に非健康なネットワークからは心疾患の原因と考えられる危険因子間の関係がわかる。

4. まとめと今後の課題

我々は項目間の関係を見出すことのできる ECM 手法の提案を行った。本手法から得られた結果に考察を加えることで、健康者と非健康者との差異を見出すことが可能になる。今後医療分野の専門家に得られた結果を分析してもらう事で本手法の有用性の確認をしたい。さらに、本研究では ECM 手法による実験だけを行ったが、医療分野で多く用いられているロジスティック回帰分析との比較を行いたい。

参考文献

[1] ECML/PKDD Discovery Challenge 2004: . http://euromise.vse.cz/challenge2004.

[2] Jaulent, M.-C., Colombet, I., Degoulet, P. and Chatellier, G.: Logistic regression model: conditions required for stability of prediction, *American Medical Informatics Association Symposium* (1999).

[3] Oka, M., Koiso, T., Meng, E. and Kato, K.: Extracting Features of Patients using the Eigen Co-occurrence Matrix Algorithm (2004). ECML/PKDD Discovery Challenge 2004 (Accepted).

[4] Oka, M., Oyama, Y., Abe, H. and Kato, K.: Anomaly Detection Using Layered Networks Based on Eigen Co-occurrence Matrix (2004). Seventh International Symposium on Recent Advances in Intrusion Detection (Accepted).