2005－MPS－57（10）
2005／12／20

PMMを用いた医療用テキストの分類

今村　裕一　　木戸　尚治　　庄野　逸
山口大学大学院理工学研究科

概要

近年，様々なテキストデータから情報を抽出し活用する研究は盛んに行われつつあり，Web 上でのデータ検索などの分野は飛躍的に発展しつつある．一方，医療機関においても設備の電子化により，カルテ等のデータが電子化され資産として蓄えられつつある．本研究では，この医療データを有効に活用するための基礎研究として，カルテに含まれる所見と呼ばれるテキストデータを幾種類かのクラスに分類することを試みた．具体的な手法としては テキストデータを単語の集まりとみなしデータ化を行なった上で，パラメトリック混合モデルによって，肺部の所見データを腫瘍，炎症，その他という 3 種類のクラスへの分類を行った．

Medical treatments classification using PMM

Imamrua Yuichi　　Kido Shoji　　Shouno Hayaru
Graduate School of Science and Engineering, Yamaguchi University

Abstract

In recent years, information mining technique for text data becomes hot topic, especially for data in the Internet web site, and a lot of methods are developed. On the other hand, a lot of clinical record becomes to be saved with computerized media in a lot medical facilities. In this research, to investigate whether those text-mining techniques are efficient for such medical treatment data, we have developed a classification application using the parametric mixture model(PMM).PMM, that assumes text data consists of “bag of words” and belongs to several given topics, infers which topics the text data are belonged to. Our medical treatment data, which describes about diseases of lungs, can be regarded as 3 types of topics, that is, “tumor”, “inflammation”, and “the other”. We tried to classify those medical treatment data.

1　はじめに

1.1　背景と目的

近年，医療機関の設備の充実により，様々なデータが電子化された媒体として記録・保存されつつある．このような蓄積された症例データから類似症例を検索できるようにすることは，医師が患者の症状に対して診断を下す際の有益な判断材料になり得る．しかしながらデータの蓄積量の肥大化により，人手による検索は困難になりつつある．

このような背景の下，本研究では電子カルテに含まれる所見や診断名と呼ばれるテキストデータに着目し，過去の類似症例検索システムの構築を試みた．所見とは，医師の診断結果や添付されている CT や MRI 画像等に対する判断や注釈，病変の経過報告などが記録されているテキストデータのことであり，診断名とは所見を通じての診断結果が記録されているテキストデータのことである．本研究では特に肺部を診断した所見をデータとして分類を試みた．肺部における所見は，主に 3 種類のトピックに分類可能であることが医師の経験により知られている．これらのトピックは “腫瘍”, “炎症”, “その他” である．所見はいずれか一つ，もしくは複数のトピックに属していると考えられている．複数のトピックに分類できることを許容しているので，この場合文書は $7(=2^3-1)$ 種類の文書に分類できることが考えられる．このような分類は文書をトピック毎に分類していると捉えることが可能であり，本研究では所見データを PMM(Parametric Mixture Model: パラメトリック混合モデル) によって同様な分離が可能かどうかを確認した [1].

PMM とはテキストのトピックに着目しテキストを単語の集合 (Bag of Words) として扱ったテキスト分類の手法である．PMM は一つのテキストが一つのトピックだけでなく，数種類のトピック（多重トピック）に分類することができる．上田らは，PMM を用いて Web 上のテキストの分類を行い良好な結果を得ている．

2　手法と対象データ

2.1　テキストの分類の流れ

ここで　PMM による識別手順についての説明を行う．まず，トピックが既知のテキスト群を用意する．次

にテキスト群の単語の出現頻度をテキスト毎に計算する (2.2.1 節参照)．次に，単語の出現頻度を用いて母数 Θ を学習させる (2.2.2 節参照)．学習させた母数 Θ を用い PMM 識別器を作成する (2.2.3 節参照)．トピックの推定ではトピックを推定したいテキストの単語の出現頻度を計算し，PMM 識別器を用いトピックの推定を行う．

2.2 PMM 識別器の作成

2.2.1 単語の出現頻度の計算

PMM 識別器を作成するためにはテキスト中の単語の頻度と必要である．ここでは以下のような所見を例に考えてみる．

> 呼吸しながらの撮像ですので呼気時に scan された部位は濃度が上昇して見えますが、吸気時に撮られた画像では濃度上昇はなく、少なくともびまん性に GGO があるのではないようです。右下葉背側 S10 に focal consolidation があり、通常は bacterial or fungal infection を疑う所見です。BOOP,leukemic infiltration も全く否定はできません。前回みられた air trapping は今回はあまり目立ちません。心拡大や胸水はありません。

但し，所見に含まれる下線部は筆者らによるものである．上記例の下線部のように所見には英単語・英略語が度々使用されている．例えば，下線部の “leukemia infiltration” という語と “白血病性浸潤” という語があったとする．このままではこの二つの語は別の単語として処理される．しかし，前者を日本語に翻訳すると “白血病性浸潤” となりこれらの単語は同じ意味を持つ．従って，予め用意しておいた英単語・英略語データベースを用い英語部分を翻訳する [2] [3]．例えば，「BOOP,leukemiinfiltration も全く否定はできません」は「器質化肺炎を伴う閉塞性細気管支炎, 白血病性浸潤 も全く否定はできません」となる．なお，英語・英略語データベースは本研究で使用するデータの英単語・英略語を中心に登録している．

医療用の所見の場合，ある症状が何の病変に由来するものかだけでなく，同様の症状をひきおこす病変を否定する表現が多様される傾向がある．このようなテキストを Bag of Words で扱った場合，否定的表現と肯定的表現とが等価に扱われ，注目したい表現の出現頻度が相対的に減少するため，分類を行う際に悪影響を及ぼすことがある．例えば提示した所見例の波線で示した部分「心拡大や胸水はありません」には，“心拡大” や “胸水” などが単語として 1 回づつ出現しているが，同じ単語が同じ回数だけ出現する「心拡大や胸水が認められます」では大分意味が異なる．従って，ここでは単語に新たな属性として「肯定」・「推定」・「否定」の 3 種類を付与し別の単語として処理する．

以上，二つの前処理を行った後，単語の頻度の計算を行う．単語の頻度の計算には予め用意しておいた単語データベースに登録されている単語のみをカウントする．なお，単語データベースは辞書，Web などから集められた人体の部位又は病名の単語が登録している [4]．

2.2.2 PMM による母数の学習

PMM はテキストが何らかのトピック (ここでは変数 Θ で表す) に属するものと仮定した上で，トピックに所属する確率を推定する手法である [1]．N 個のトピックと単語頻度が既知のテキスト群 $\mathcal{D}$ を用意する．Bag of Word として扱った場合テキストデータは単語の出現ベクトルで表わすことが出来る．n 番目のテキストデータ d_n の単語頻度を表現するベクトルを $\boldsymbol{x}_n = (x_{n,1}, \cdots x_{n,V})$ と表わす．V は全てのテキストで検出された語群の総数に一致する．一方，トピックも同様に $\boldsymbol{y}_n = (y_{n,1}, \cdots x_{n,L})$ と表し，第 l トピックに属す (属さない) 時，1(0) の値をとる 2 値とした．L は総トピック数，V は語群総数で既知とである．

次にテキスト d_n が第 l トピックに属する経験確率 $h_l(\boldsymbol{y}_n)$ とする．

$$h_l(\boldsymbol{y}_n) = \frac{y_{n,l}}{\sum_{l'=1}^{L} y_{n,l'}} \tag{1}$$

となる．l 番目のトピックに i 番目の単語が出現する確率を $\theta_{l,i}$ $(l = 1, 2, \cdots, L)$, $(i = 1, 2, \cdots, V)$, $(\theta_{l,i} \geqq 0, \sum_{i=1}^{V} \theta_{l,i} = 1)$ と定義すると PMM は次式で表すことができる．

$$\begin{aligned} p(d_n|\boldsymbol{y}, \Theta) \propto \prod_{i=1}^{V} \Phi^{x_{n,i}} &= \prod_{i=1}^{V} \left(\sum_{l=1}^{L} h_l(\boldsymbol{y}) \theta_{l,i} \right)^{x_{n,i}} \\ &= \prod_{i=1}^{V} \left(\frac{\sum_{l=1}^{L} \boldsymbol{y_l} \theta_{l,i}}{\sum_{l'=1}^{L} \boldsymbol{y_{l'}}} \right)^{x_{n,i}} \end{aligned} \tag{2}$$

但し，$\Theta = (\theta_{1,1}, \theta_{1,2} \cdots, \theta_{1,V} \cdots, \theta_{L,V})$

PMM は全ての可能な多重トピッククラスを L 個のパラメータベクトルで完備かつ効率的に表現されている．

以上より，トピックが既知のテキスト郡 $\mathcal{D}$ が与えられた際，未知パラメータ Θ は事後分布 $p(\Phi|\mathcal{D})$ を最大化するパラメータとして求められる．Θ に関する最大化問題は解析的に解くことができず，逐次反復法で導出する．

第 t ステップでのパラメータの推定値を $\Theta^{(t)}$(既知) として，$\Theta^{(t+1)}$ の推定値を導出する．導出した結果，Θ のパラメータ学習の更新式は以下のようになる．

$$\theta_{l,i}^{(t+1)} = \frac{\sum_{n=1}^{N} x_{n,i} g_{l,i}^{n}(\theta^{(t)}) + 1}{\sum_{i=1}^{V} \sum_{n=1}^{N} x_{n,i} g_{l,i}^{n}(\theta^{(t)}) + V} \quad (t = 0, 1, 2. \cdots)$$

$$\text{但し，} \quad g_{l,i}^{n}(\theta) = \frac{h_l(\boldsymbol{y}_n)\theta_{l,i}}{\sum_{l'=1}^{L} h_{l'}(\boldsymbol{y}_n)\theta_{l',i}} \tag{3}$$

Θ のパラメータ推定アルゴリズムは，初期値の如何にかかわらず大域的最適解に収束する．本研究では計算の終了条件[1]を

$$|\theta_{l,i}^{(t+1)} - \theta_{l,i}^{(t)}| = \varepsilon \tag{4}$$

とした．なお，ε は収束判定閾値で 10^{-5} とした．

[1] 本研究では t=25 ほどで収束した

2.2.3 PMM 識別器の作成

トピックが未知かつ単語の頻度が既知のテキストのトピックの推定では Θ の推定問題と双対構造をなす．即ち，Θ のパラメータ推定では，$\boldsymbol{h}$ を既知，Θ を未知としていたが，ここでは，Θ を既知として $\boldsymbol{h}$ を未知とした問題に置き換える．従って，Θ のパラメータ学習と同様な導出により，以下のパラメータ更新式を $\boldsymbol{h}$ を得る．

$$h_l^{(t+1)} = \frac{\sum_{i=1}^{V} x_i g_{l,i}(\boldsymbol{h}^{(t)}) + 1}{\sum_{i=1}^{V}\sum_{l=1}^{L} x_i g_{l,i}(\boldsymbol{h}^{(t)}) + V} \quad (t = 0, 1, 2. \cdots) \tag{5}$$

となり，PMM 識別器を得る．また，$\boldsymbol{h}$ のパラメータ推定アルゴリズムは，初期値の如何にかかわらず大域的最適解に収束する．

3 実験と結果

3.1 対象データ

本研究で使用するデータは，山口大学医学部附属病院より提供を受けたデータである．肺野について記述さている 254 症例を診断名に着目し，所見を 3 種類のトピックを用い 7 通りに分類した．使用するデータの基本統計量を表 1 に示す．単語総数は単語データベースと合致した所見の単語数である．また，トピックの多重度は 22.4%であった．すなわち，77.6%の所見が単一トピックに属すといえる．

表 1: 基本統計量

単語総数	平均単語数	多重度 (%)
1426	29.0	22.4

3.2 実験

分類した 254 症例の所見の中から 1 つの所見を取り出し，トピックが未知，単語頻度が既知としサンプルデータとした．そして残りの 253 症例の所見のトピックと単語頻度が既知とし訓練データとした．訓練データを用い PMM 識別器を作成しサンプルデータのトピックの推定を行った．この操作を 254 症例の所見全てに対してそれぞれ行った (Leave one out 法)．

分類の結果は，複数選択を許した上でのトピックに対する所属確率となるので，大まかにわけると前述の通り 7 通りになる．ここでは X 軸に “腫瘍”の推定値, Y 軸，Z 軸それぞれに “炎症”と “その他”の推定値を割り当てるものとし表示を行っている．ただし各推定値は確率であるため全てを足すと 1 になることから $X + Y + Z = 1$ の平面上の三角形内部の点としてプロットできることになる (図 6 参照)．図 1〜5 はこの各データの推定値をこの三角形内部の点と指定プロットしたものであり，頂点に近いデータほど単一のトピックに含まれているデータとして推定されていることになる．その反対に所属が不明瞭なデータほど三角形内部の点として表現されることになる．図 1 は腫瘍のみに所属していると考えられているテキストを分類した結果であり，三角形の左上の頂点にクラスタがあるのが見てとれる．なお，図中の塗りつぶしたシンボルとその周囲に広がる円は，それぞれデータの平均座標と分散値を表している．同様に図 2 が炎症のみに所属しているデータ，図 3 はその他のみに所属しているデータを分類した結果で，図 1 と同様に各頂点にクラスタを作る傾向にある．図 4 と図 5 は複数トピックに所属しているデータを分類した結果でありトポロジカルにはほぼ正しい位置にクラスタを作っているように見える．なお，図 4 の逆三角形は “腫瘍と炎症”に，五角形は “腫瘍とその他”に所属しているデータであり，図 5 の菱形は “炎症とその他”に，六角形は “腫瘍と炎症とその他”に所属しているデータである．

3.3 医師による分類との比較

PMM 識別器が人手によるトピックの推定と，どの程度，差があるのか比較した．254 症例の所見の中から 22 症例をランダムに取り出し，医師が手動で 22 症例を “腫瘍”，“炎症”，“その他”に分類し三角形内部にマークした．マークした点より，“腫瘍”，“炎症”，“その他”の推定値を算出した．医師による分類結果と PMM 識別器で算出した推定値の差をユークリッド距離[2]で求めた．結果を図 7 に示す．横軸は距離であり，縦軸は所見数を表す．

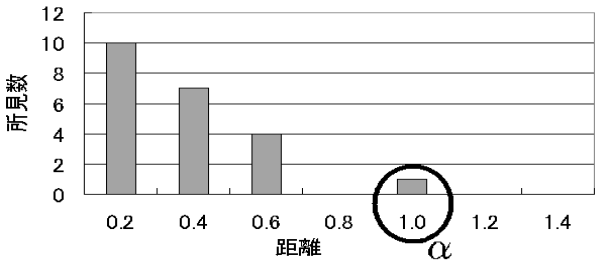

図 7: PMM 識別器と医師による分類の差のヒストグラム
図中の丸 (α) は 4 章参照

4 考察

図 1〜5 を見る限り，ほぼ正しい位置にクラスタを作っているのがわかるが分散値を見るとかなりの範囲にひろがっているのも見て取れる．したがって検索結果を提供する場合には，検索範囲をある程度に絞り込む程度にとどめておく方が有益と考えられる．ただし，個々の例を見ていくと明らかに分類に失敗している所見も見られる．

失敗した所見の多くは，診断名のトピックと異なるトピックに関して書かれている所見であった．例えば診断名が「肺癌」や「肺癌疑い」の場合，互いにトピックは “腫瘍”である．しかし，前者は “腫瘍”に関してのみ書かれていることが多かったが，後者は他の病気の可能性を考えるため他のトピックに関しても書かれて

[2]最小値は 0，最大値は $\sqrt{2}$ である．

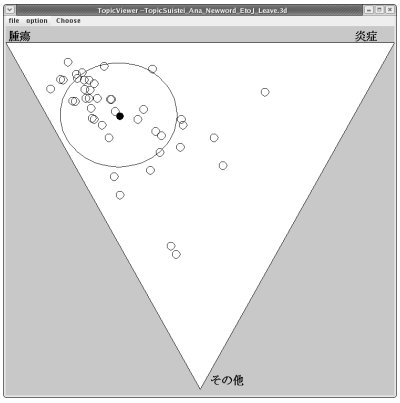

図 1: 腫瘍

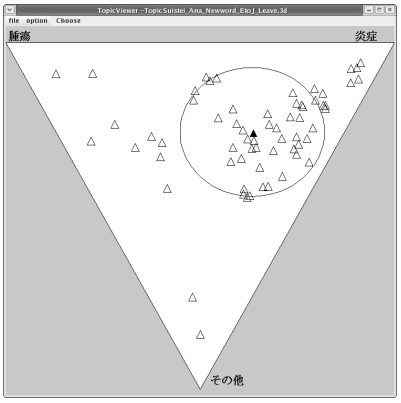

図 2: 炎症

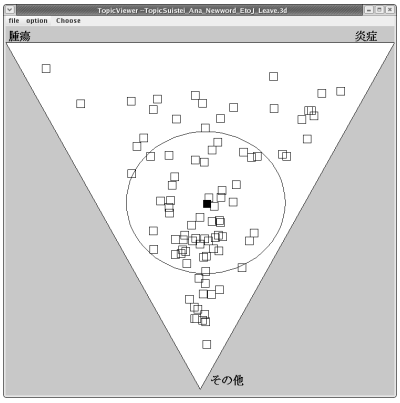

図 3: その他

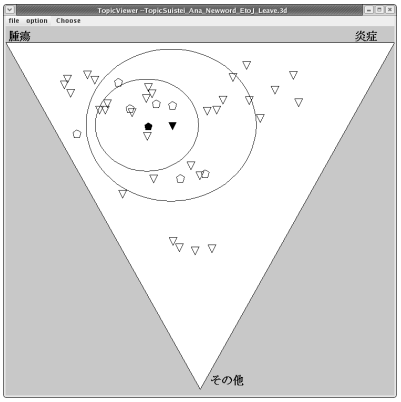

図 4: 腫瘍と炎症 (逆三角形)，腫瘍とその他 (五角形)

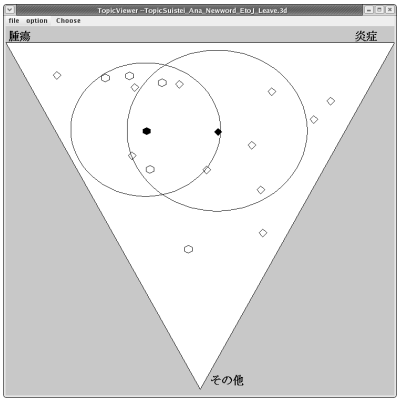

図 5: 炎症とその他 (菱形), 腫瘍と炎症とその他 (六角形)

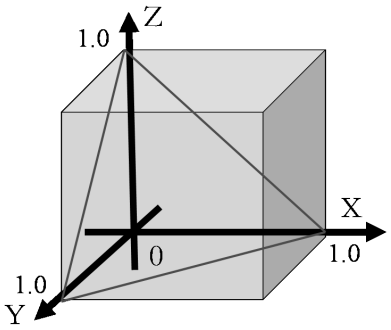

図 6: 図 1〜5 の概念図

いることが多かった．つまり，診断が定まらない所見は他のトピックに関して書かれている傾向にあった．従って，訓練データは「診断名と所見のトピックが一致しているもの」，つまり，診断がはっきりしているものを選ぶべきであった．

次に識別実験の結果 (図 7) を見る限りでは，医師の分類結果と PMM 識別器の分類結果はその距離が比較的小さいところに集まっている事から同じ傾向を示していると考えることができる．しかし，かなり異なった分類を行う場合もある (図 7 中の丸)．この所見を所見 α とする．所見 α の推定値 (表 2 参照) は PMM 識別器では腫瘍が最も高く，医師による分類では炎症が最も高い．これは所見 α が CT や MRI 等の画像に対する所見だけではなく，手術に関する報告が書かれているためだと考えられる．実際，所見 α の出現回数の多い上位 10 個の単語のうち，5 つは腫瘍に対する出現確率が最も高かった．

PMM では単語の出現頻度に着目し，多く出現した単語の出現確率が大きく反映される．しかし，医師による分類の場合，出現頻度は関係なく一語一語の意味を考慮し分類する．従って，所見 α のように診断に余り意味をなさない単語が多く含まれる場合，PMM 識別器の分類と医師との分類では大きな差が生じてしまう．一方，診断がはっきりしている所見において，差はほとんどなかった．

また，医師の所見の分類には主観が入ることも考えられ，客観性を上げるため複数の医師による所見の分類などの対策も必要であると考えられる．

今後，訓練データの数を増やし，かつ診断が確定していない所見は訓練データからはずして PMM 識別器を作成し，どの程度識別率があがるか試みる予定である．

表 2: 最も差が大きかった所見 α のトピック推定値

PMM 識別器			医師		
腫瘍	炎症	その他	腫瘍	炎症	その他
0.87	0.07	0.06	0.23	0.53	0.24

参考文献

[1] 上田修功，斉藤和巳，”類似テキスト検索のための多重トピックステキストモデル ”，情報処理学会研究報告（情報処理学会　2003-5-9)，p17-20

[2] 真柄正直，真柄婦美，”英和医語中辞典 ”，文光堂

[3] 医学生のための医学略語集, http://www.med-pearls.com/text/abb/abb.htm

[4] 診療報酬情報提供サービス, http://202.214.127.149/