大規模テキストデータの分類体系化のための機械学習に基づく半自動 KJ 法の提案

1X08C070-9　下村　良
指導教員　後藤正幸

1　研究背景と目的

今日，多くの企業が業務上で発生するテキスト情報をデータとして大量に蓄積している．蓄積されたデータには，各個人が保有する業務ノウハウ等の暗黙知が含まれている可能性があり，有効活用が望まれる．有効活用の一案として，データの構造化が考えられる．蓄積されたデータを体系立てて整理することで，その構造全体を俯瞰することができ，企業活動における様々な場面での活用が期待できる．

データを体系化する手法として KJ 法 [1] が知られている．KJ 法とは，意味的に似たデータ同士をグルーピングし，いくつかのカテゴリを作ることでデータの全体像を捉えることを目的とした手法である．しかし KJ 法では全てのデータを人手で仕分けする必要があるため，大規模なデータに対して適用することは難しい．

一方，大規模なテキストデータを扱う技術として自動文書分類がある．自動文書分類とは，分類整理された少量のデータを基に，残りのデータを自動的に分類する技術のことである．この技術を援用することにより，KJ 法の仕分けによる負荷の緩和が期待できる．

そこで本研究では上記 2 つの手法に着目し，これらを組み合わせることで大規模データを効率良く整理する手法を提案する．具体的には，比較的少数の部分データ集合を人手で分類し，このデータを基に分類器の学習を行って残りの大量のデータを自動的に分類することで効率的にデータを体系化する方法を構成する．また，ソフトウェア開発の一端を担う企業（以下，X 社とする）が保有する大規模な実データに対して提案手法を適用し，その有効性を示す．

2　テンプレートマッチング

テンプレートマッチングとは，ベクトル空間モデルを用いたパターン認識手法である．事前にカテゴリ毎の代表となる重心（テンプレート）を求めておき，対象データとの距離が最も近い重心を持つカテゴリにデータを分類する．

カテゴリ数 N を既知とし，カテゴリの集合を $\mathcal{C} = \{C_1, C_2, \cdots, C_N\}$ とする．カテゴリが既知の文書数を D, 文書集合を $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_D\}$ とし，$\mathcal{X}$ 内の文書に含まれる総異なり単語数を J, 単語集合を $\mathcal{W} = \{w_1, w_2, \cdots, w_J\}$ とする．このとき，ある文書 $\boldsymbol{x}_i$ は，$\mathcal{W}$ で構成されるベクトル空間によって表され，$\boldsymbol{x}_i = (t_{i1}, t_{i2}, \cdots, t_{iJ})$ となる．ただし，t_{ij} は i 番目の文書における j 番目の単語の出現頻度である．また，カテゴリ C_n の重心は J 次元ベクトル $\boldsymbol{b}_n$ で表され，(1) 式で与えられる．

$$\boldsymbol{b}_n = \frac{1}{|C_n|} \sum_{\boldsymbol{x}_i \in C_n} \boldsymbol{x}_i \tag{1}$$

ただし，$|C_n|$ はカテゴリ C_n に属する文書数である．このとき，カテゴリが未知の新たな入力文書を $\boldsymbol{x}$ とすると，$\boldsymbol{x}$ の所属カテゴリを (2) 式で推定する．

$$\hat{C} = \arg\min_{C_n} d(\boldsymbol{x}, \boldsymbol{b}_n) \tag{2}$$

なお，$d(\boldsymbol{x}, \boldsymbol{b}_n)$ は $\boldsymbol{x}$ と $\boldsymbol{b}_n$ の距離である．今回は距離尺度としてコサイン尺度を基にした (3) 式を用いた．

$$d(\boldsymbol{x}, \boldsymbol{b}_n) = 1 - \frac{\boldsymbol{x} \cdot \boldsymbol{b}_n}{||\boldsymbol{x}||\ ||\boldsymbol{b}_n||} \tag{3}$$

3　提案手法

3.1　問題設定

本研究で解析の対象とするデータは，

- 大規模なテキストデータであること
- 分析者が分類の指針とカテゴリの種類を与えること

を前提としている．前者は，人手では全てを仕分けできないほどの膨大なデータ量であることを意味する．後者は，対象のデータが属しえるカテゴリの意味づけが分析者により人手で行われるものであることを指す．これらの前提は，テキストマイニング手法の一つであるクラスタリングに用いられるものとほぼ同様である．しかし, クラスタリングでは, クラスタの意味づけが可能か否かに関わらず，データの集まりが受動的に作成されるのに対し，提案手法では，カテゴリの意味や名前が能動的に定められるという点でこれら 2 つの手法は明確に異なっている．

3.2　テンプレートマッチングを用いた半自動 KJ 法

提案手法のフローを図 1 に示す．

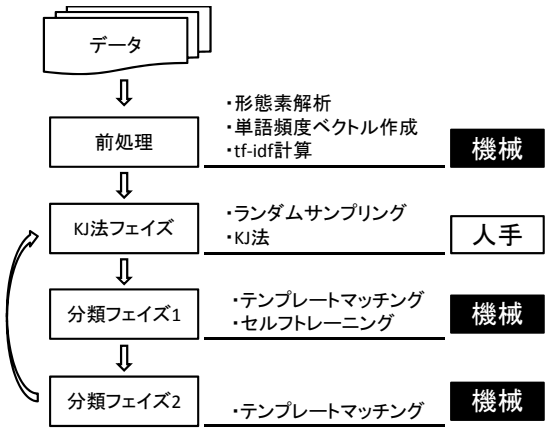

図 1. 提案手法のフロー

本手法は，始めに少量のデータを人手で分類し，その分類方針を学習させた分類器を用いて残りの大量のデータを自動分類することで効率的な体系化を目指している．この作業を繰り返すことにより，人手の分類負荷を軽減することができ，大規模データの分類整理が可能となる．

ただし，分類フェイズでは，カテゴリごとに学習データ数にばらつきがある上，カテゴリ数が多い場合には 1 カテゴリ当たりの学習データ数が減少するため，一部のカテゴリの学習データ数が極端に少なくなる恐れがある．そこで，半教師つき学習のセルフトレーニング [2] の概念を用い，分類フェイズを二段階に分けて行う．初めの分類フェイズ 1 では学習データ数の増加を目的とし，学習データ数が一定数以上になるように少量のデータの分類を行う．その後，分類フェイズ 2 にて，残りの大量のデータの分類を行う．

3.3　アルゴリズム

カテゴリ情報が付与されたデータをラベルありデータ，カテゴリ情報が付与されていないデータをラベルなしデータと

表 1．分類結果（一部抜粋）

大カテゴリ	カテゴリ集合	通し番号	例	度数	カテゴリ検出時の繰り返し数k
テキスト関連	感性の問題	1	「○○○」は「×××」と表記された方が文章が自然かと思われます。	1162	1
	表示の際の問題	2	上記二行目行頭「。」は行頭禁則に抵触する可能性があります。	1899	1
	複数表現	3	「○○○」と「×××」で相違しております。	1925	1
	表記誤り	4	上記「○○○」は「×××」と表記されるのが適切かと思われます。	8650	1
	余計な空白	5	上記「○○○！」の感嘆符の後に一文字分の空白が空いておりません。	86	2
	話の流れとの不一致	6	「○○」の説明テキストにおいて「○○」の設定ポジションがDFの為、「×××」との表記は不自然かと思われます。	80	3
	ロジックの不具合	7	フィールド上の○○○のあるところで、敵とエンカウント致します。	8051	1
		8	フィールド南西部にある町に入[illegible]ます。	204	2
その他	通信時の不具合	[illegible]	[illegible]て、バックアップ致します。	[illegible]	[illegible]
	処理落ち	27	「○○○」において、プレイ中に処理落ちが発生しております。	99	3
	倫理的問題	28	「○○」の台詞テキストに、「×××」とありますが、上記赤字部は性的なものを彷彿させますが、倫理的に問題ありませんでしょうか。	61	3
	説明不足	29	プレイヤーが「○○○」の存在に気付いても、接触せずに通過してしまう恐れがあるものと思われます。「○○○」の説明においても、一度触れる必要がある旨を記載された方がよりユーザーフレンドリーかと思われます。	25	4

する．カテゴリの集合を $\mathcal{C}$ とし，ラベルありデータの集合を $\mathcal{R}$，ラベルなしデータの集合を $\mathcal{U}$ とする．提案手法のアルゴリズムを以下に示す．

step1) $\mathcal{C}=\phi, \mathcal{R}=\phi, \mathcal{U}$ を全データとする．k=1 とする．

step2) パラメータ $\alpha_k(0<\alpha_k<1)$ を定め，$q_k=\lceil \alpha_k|\mathcal{U}| \rceil$ とする ($|\mathcal{U}|$ は $\mathcal{U}$ の要素数，$\lceil \cdot \rceil$ は適当な整数値に切り上げる操作を表す)．q_k 件のデータを無作為に抽出し，その集合を $\mathcal{L}_k$ とする．$\mathcal{U}\backslash\mathcal{L}_k$ を新たな $\mathcal{U}$ とする (ランダムサンプリング)．

step3) 抽出した q_k 件のデータのうち直感的に似ていると思うものをグルーピングし，各グループにタイトルを付ける．本ステップでタイトルが付与されたグループの集合を $\mathcal{G}_k$ とする．$\mathcal{C}\cup\mathcal{G}_k$ を新たな $\mathcal{C}$ とする (KJ 法)．

step4) $\mathcal{G}_k$ をカテゴリ集合とみなし，$\mathcal{L}_k$ にカテゴリ情報を付与したデータを学習データとして分類器に学習させ，$\mathcal{U}$ を $\mathcal{G}_k$ のいずれかのカテゴリに分類する (テンプレートマッチング)．

step5) カテゴリ毎に閾値 β_n を設定し，$\mathcal{U}$ の各データと分類されたカテゴリの重心との距離が β_n 未満のデータを正しく分類されたとみなし，その集合を $\mathcal{T}$ とする．$\mathcal{L}_k\cup\mathcal{T}$ を新たな $\mathcal{L}_k$ とし，$\mathcal{U}\backslash\mathcal{T}$ を新たな $\mathcal{U}$ とする．

step6) 再び，step4 と step5 を繰り返す．

step7) $\mathcal{R}\cup\mathcal{L}_k$ を新たな $\mathcal{R}$ とする．$\mathcal{U}=\phi$，もしくは $|\mathcal{L}_k|<3\times\alpha_k$ ならば $\mathcal{R}$ と $\mathcal{C}$ を出力し，アルゴリズムを終了する．それ以外は $k+1$ を新たな k として step2 に戻る．

□

なお各 k において，最初の step4 と step5 が図 1 の分類フェイズ 1 に当たり，step6 が分類フェイズ 2 に当たる．

4 実験

提案手法の有効性を示すため，X 社が保有する実データの体系化を目的とした実験を行った．

4.1 実験条件

テストデータには，X 社が試験工程で検出したソフトウェアの不具合 59,085 件のデータを用いた．このデータには各不具合の発生手順や症状がテキスト形式で記載されており，不具合 1 件分のデータを 1 件のテストデータとみなして実験を行った．また，パラメータ α_k と β_n は，適宜，実験的に適当な値を設定した．効率性の評価は，各繰り返し数 k における KJ 法に用いるラベルなしデータ数 q_k，ラベルありデータ数 $|\mathcal{R}|$，ラベルなしデータ数 $|\mathcal{U}|$ の値を比較し，人手の作業量をどの程度軽減できたかを検証することで行う．

4.2 実験結果

提案手法の適用結果を表 1, 2 に示す．表 1 の度数はそのカテゴリに属するテストデータ数，検出ステップ数はそのカテゴリが発見された際の本手法の繰り返し数 k を表す．なお抽出されたカテゴリには類似したものが存在したため，それらをグルーピングし，大カテゴリとした．表 1 より未整理のテストデータから 29 のカテゴリを抽出できたことがわかる．

表 2 は，k 回目の繰り返しにおける各データ数を示す．計 5 回の繰り返しによって 48,000 件以上のテストデータにカテゴリ情報が付与された．これは全データの約 81.2 %に当たる．一方で実際に人手で分類したテストデータは 1,000 件に留まっており，全体の 1.7 %程度である．これは人手の作業量が膨大であるという KJ 法の欠点を，テンプレートマッチングによって補完できた結果と考えられる．

表 2．繰り返し数 k の増加に伴うデータ数の推移

繰り返し数k	1	2	3	4	5
KJ法に用いるデータ数q_k	300	200	200	100	100
ラベルありデータ数$\lvert\mathcal{R}\rvert$	36739	45111	46641	47999	48245
ラベルなしデータ数$\lvert\mathcal{U}\rvert$	22346	13974	12444	11086	10840

5 考察

表 1 から，度数の高いカテゴリほど早い繰り返し数 k で検出され，k の増加に伴って度数の低いカテゴリが検出される傾向にあることが窺える．これは，ラベルありデータに類似したテストデータの自動分類が有効に働いた結果と考えられる．また「ロジックの不具合」カテゴリのように，他と比べて度数が極端に高いカテゴリができている．こういったカテゴリは範囲の広い抽象的なカテゴリであり，再帰的に提案手法を適用することで，更なる細分化が可能であると思われる．

表 2 では，k の増加に伴う $|\mathcal{R}|$ の増加量が単調減少している．$\mathcal{U}$ には，類似性が低く自動分類が難しいテストデータが多く残っており，k を更に増やしても効率的な体系化は難しいものと思われる．こうしたデータは度数の低い一般的でないデータであると考えられ，別途，新たな分類体系化の仕組みを検討する必要がある．

6 まとめと今後の課題

本研究では，KJ 法による人手の負荷をテンプレートマッチングで軽減した手法を提案し，大規模データに対しての適用を試みた．その結果，全体の 1.7 %程度のテストデータを手作業で整理することで，81.2 %のテストデータにカテゴリ情報が付与され，29 のカテゴリを抽出することができた．これにより，データの効率的な体系化を目指す本手法の有効性が示されたと言える．

しかし，提案手法で抽出されるカテゴリは属人的なものであり，作業者によってアウトプットが異なるという性質がある．今後は，アウトプットの良し悪しを判断する評価手法を検討する必要がある．

参考文献

[1] 川喜田二郎, “KJ 法—混沌をして語らしめる,” 中央公論社, pp. 431-438, 1986 年 11 月.

[2] M. Seeger, “Learning with labeled and unlabeled data,” *Technical report*, University of Edinburgh, Dec. 2002.