名前同定のためのSVM特徴素の抽出と適用※

港　真人†　　相澤　彰子‡

日立工業専門学院†　　国立情報学研究所‡

1　はじめに

大規模データベースにおいては，入力誤りやプログラムによる変換誤りなどの人的ミスや姓名・住所・価格の時間変化によるデータの劣化は避けて通ることができない問題である．これらの要因によって，一ヶ月間に約2%のデータが劣化するとの報告もあり[1]，経済的な損失をはじめとして，その影響は甚大である．

不整合なデータを検出してデータベースの品質を維持するためには，値が完全には一致しないが同一の実体を参照しているレコードどうしを抽出する「レコード同定」の技術が重要である[2]．本稿では，特に社会的なニーズが高い人物の同定に焦点をあて，レコード同定の性能向上を目指す．人物の同定においては，主に名前が判断においての中心的な役割を果たすが，曖昧性解消のためにはその他の属性（住所や生年月日，論文であれば共著者や所属等）も必須である．そこで，本稿では他の属性情報に基づき確信度高く同一だとみなされる氏名ペアを抽出することで，異体字や混同しやすい漢字などを判別器の特徴素として自動獲得し，人物同定精度を高める手法を提案し，実験により有効性を確認する．

2　提案手法

多くのデータベースでは，レコードの属性に冗長性を持たせることで，レコードを信頼度高く識別することが可能になっている．たとえば，郵便番号と住所，製品番号と製品名などである．本稿では，この冗長性を利用して共訓練の枠組[3]の適用を試みる．ただし，ここで想定する「事例」は単独のレコードではなく，2つのレコードの組み合わせに対して2値のラベルを割り当てたものである．事例の大多数が負例で，正例どうしの重なりも大きくないことを考慮すると，自動生成したラベルを追加するだけでは，求められる同定性能を達成することは難しい．

そこで共訓練を，ラベルの獲得ではなく，手がかりとして有効な特徴素の獲得のために用いる．

いま，同定対象とするレコードの集合を R (=$\{r_1, ..., r_m\}$)，レコードの属性集合を A (=$\{a_1, ..., a_n\}$)とする．任意の2つのレコード $r_1, r_2 \in R$ に対して，$f(r_1, r_2)=L$ (L=1のとき「一致する」，L=0 のとき「一致しない」) なる2値ラベルを割り当てる関数 fを同定関数と呼ぶ．本稿で想定する同定関数は，各属性ごとに定義される特徴素集合 $\lambda(a_1), ..., \lambda(a_p)$を入力とする分類器で，実装ではSVMを用いた．$\lambda(a_i)$はたとえば“名前の編集距離”などを含む．提案手法では，以下の手順で特徴素の抽出を行う．

① 同定の手がかりとして十分な属性部分集合 $I \subset A$を人手で選択する．

② 特徴素集合$\{\lambda(a_i) \mid a_i \in I\}$ を用いて，確信度高く同じであると判断されるレコード対を獲得する．

③ 得られたレコード対を用いて，注目する属性 $a_j \in A$ の特徴素候補を評価し，有効な特徴素を$\lambda(a_j)$に追加する．

これによってデータベース固有の入力誤りなどを特徴素として容易に取り込むことが可能になると考えられる．

3　論文著者の同定問題への適用

日本人の名前を判定する際の問題として，異体字や混同しやすい漢字などがある．これらは入力ミスや記録媒体の移し替え等によって間違う可能性が高く，個々のデータベースに依存する傾向が強いと考えられる．そこで，本研究では学術論文データベースを利用し，この間違われやすい異体字や混同しやすい漢字を自動抽出して同定性能の向上を試みた．

具体的には，論文著者レコードの属性集合を{“和文氏名”，“英文氏名”，“和文所属”，“英文所属”，“共著者の和文氏名”}として，同定に十分な属性集合として{“和文氏名”, “共著者の和文氏名”}を選んだ．そして，”和文氏名が1文字違い”かつ”共著者名が3名一致する”著者レコード同士を同一ペアとして獲得した．次に，この著者

※「A Method for Extracting SVM Features for Name Identification and Its Application.」

†「Masato Minato: Hitachi Technical Academy」

‡「Akiko Aizawa: National Institute of Informatics」

レコードの“和文氏名”の異なる1文字ペアの頻度を調べ，出現回数が上位のペアを新たな特徴素として追加した．最後に，別途準備した人手判定による正解データを用いてSVMを再学習させた．

4 実験

4.1 実験について

実験では，約1億2571万件の論文ペアを対象に，約2万件の文字ペアを抽出した．抽出した文字ペアの一部を表1に示す．

表1　自動抽出した異なる1文字ペア

頻度(件)	文字ペア
234996	斉 斎
206012	廣 広
95101	郎 朗
64332	裕 祐
58488	巳 己
⋮	
4883	泰 =

表1を見ると，(斉,斎)や(廣,広)は異体字であり，(郎,朗)や(裕,祐)，(巳,己)等は恐らく入力の際に間違われたであろう漢字，いわゆる混同しやすい漢字であると思われる．また，(泰,=)のように文字コード上表せない文字を=(下駄文字)とする場合もあった．

さらに，あらかじめ人手で同一著者であるかを判定した著者データペア15648件(正解データ12083件，不正解データ3565件)を用い，SVMの性能を評価した．比較のため，名前の類似度や氏名の文字列長の違い等人手で設定できる素性を特徴素とした場合(STA)と，そこに異体字，混同しやすい漢字等の動的に抽出された特徴素を追加した場合(DYN)に分け評価を行った．STAとDYNで共通に用いた特徴素を一部以下に示す．

- 氏名，所属の類似度
  編集距離やLCSをもとに計算した．
- 氏名の文字列長の違い
  一般に，氏名の長さが異なると別人だと考えられる．(“田中健太”と”田中健太郎”等)
- 英語表記でのイニシャルの一致
  英語表記の場合，名がイニシャル表記となるケースが多いため．

4.2 実験結果

SVMによる人物同定の評価結果を表2に示す．本研究では動的に抽出した特徴素の効果を調べるため，判定済データ15648件の内，和文氏名が異なるもの3024件(同一人物2265件，異なる人物759件)を用いて5分割交差検定により評価した．また，追加特徴素とした異なる1文字ペアについては抽出した約2万件の内，最も性能の良かった上位800件を用いた．SVMのカーネル関数には2次多項式を用いている．

表中のAccuracyは同定の精度，Precisionは適合率で別人物が同一人物であるとされていないか，Recallは再現率で同一人物の見落としが無いかを表す．

表2　評価結果

	Accuracy	Precision	Recall
STA	87.36	91.20	92.41
DYN	90.28	92.42	94.79

4.3 考察

評価用データは判定が困難であるもの（氏名が若干違う，氏名が同じで所属が異なる，同姓同名等）を中心に構成されているが，それでも高い同定性能を示している．特にSTAに比べてDYNでは何れの数値も高く，動的に抽出した特徴素が有効であると考えられる．しかし，DYNを追加した結果，判定ミスとなった物もあり，より効果的な抽出手法を検討する必要がある．

5 おわりに

本研究では，人物同定の為の動的なSVM特徴素の抽出手法について提案した．また，学術論文データベースに人物同定を適用し著者同定を行い，同定精度の有効性を検証した．

著者同定の結果を用いることで，研究者ネットワークの構築や，論文推薦など様々な用途への貢献が期待できると考える．

参考文献

[1] Koudas, N., Sarawagi, S., and Srivastava, D.: “Record linkage: similarity measures and algorithms,” in *Proc. of ACM SIGMOD 2006*, 802-803 (2006).

[2] Thomas N. H., Fritz J. S., and William E. W.: *Data Quality and Record Linkage Techniques*, Springer (2007).

[3] Blum, A. and Mitchell, T.: “Combining labeled and unlabeled data with co-training,” in *Proc. of COLT' 98*, 92-100 (1998).