高精度なツイート投稿位置の推定に有効なルールの抽出

宮内　天士† 　白川　真澄†† 　原　隆浩†† 　西尾章治郎††

† 大阪大学工学部　〒 565–0871 大阪府吹田市山田丘 2-1
†† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1-5
E-mail: †{miyauchi.takashi,shirakawa.masumi,hara,nishio}@ist.osaka-u.ac.jp

あらまし　位置情報付きツイートは，ローカルイベントの検出やユーザの行動予測など様々な研究に利用されている．しかし，位置情報付きツイートは全ツイートに対してごく少量しかなく，位置情報付きツイートを用いた研究ではこのリソースの少なさが問題となる．そこで本稿では，位置情報付きツイートの総量を増やすことを目的とし，ツイートに含まれるスポット名がそのツイートの投稿位置を指すようなケースを高い精度で検出する手法を提案する．提案手法では，スポット名を含む位置情報付きツイートを教師データとして，どのような特徴の組合せ（ルール）をツイートが有するときにスポット名に対応する位置情報とツイートに付与されている位置情報とが一致するかを分析することで，ツイートの投稿位置推定に有効なルールを抽出する．また，ツイートの特徴の数は膨大であるため，その組合せを効率よく探索するためのルールマイニング手法を提案する．評価実験により，線形 SVM を用いた手法と比較して，提案手法が高い適合率を達成しつつ，より多くのツイートに対して位置情報を付与できることを確認した．

キーワード　マイクロブログ，Twitter，位置推定，相関ルール

1. はじめに

マイクロブログの普及により，実世界で起こっている出来事や関心事についての情報をリアルタイムに発信することが可能になった．マイクロブログの代表的なサービスである Twitter では，ツイートと呼ばれる 140 文字までのテキストメッセージを投稿・共有できる．ツイートはリアルタイム性の高い情報を多く含んでいることから，Twitter を対象とした研究が盛んに行なわれている．また，Twitter は携帯電話やスマートフォンと連動させることにより，端末の位置情報（緯度経度情報）をツイートに付与して投稿することができる．この位置情報付きツイートを利用して，地震や祭などのローカルイベントの検出，ユーザの行動予測，各地域のホットトピック抽出など，様々な研究が行われている．しかし，位置情報付きツイートは全体に対してごく少量しかなく，位置情報付きツイートを利用する研究では，これらのリソースの少なさが問題となっている [1]．

位置情報付きツイートのリソースの少なさに対処する一般的な方法として，ジオコーディングと呼ばれる技術がある．ツイートに対するジオコーディングでは，地名やスポット名を含むツイートに対してその場所の位置情報を付与する．地名やスポット名と同じ文字列が偶然ツイートに含まれている場合や，それらがニュースの話題語として出現している場合などでは，投稿時のユーザの位置情報とは異なる位置情報を付与することになる．しかし，リアルタイムなユーザの位置情報を利用するアプリケーションや研究においては，ツイート投稿時のユーザの位置情報と付与したユーザの位置情報が同じである必要がある．

そこで筆者らの先行研究 [6] では，位置情報付きツイートの総量を増やすことを目的とし，ジオコーディングによって位置情報を付与可能なツイートのうち，ツイート投稿時のユーザの位置がスポット名の位置と一致する場合のみを，線形 SVM（線形カーネルを用いた SVM）を用いて検出する手法を提案した．この手法では，ツイート中のスポット名が指す位置情報とツイートに付いている位置情報が一致する場合，一致しない場合をそれぞれ正例，負例として，ツイートから抽出可能な様々な特徴量を学習する．学習時には位置情報付きツイートを用いるため，教師データを自動的に生成できるという利点がある．この手法により，位置情報の付いていないツイートの一部について，高い適合率でツイート投稿時のユーザの位置情報を付与できる．しかし，先行研究の手法では，線形 SVM により各特徴量に対する重みを学習できるが，特徴の組合せについては学習できない．特徴の組合せを分析することにより，高い適合率を維持したまま，さらに多くのツイートに位置情報を付与できる可能性がある．

そこで本研究では，相関ルールマイニングを用いて特徴の組合せを分析することで，高い適合率を維持しつつ，より多くのツイートに位置情報を付与することを目指す．具体的には，スポット名を含む位置情報付きツイートを教師データとして，どのような特徴の組合せ（ルール）が該当するときに，そのスポット名に対応する位置情報とツイートに付与されている位置情報とが一致するかを，可能な限り網羅的に探索する．ここで，ツイートの特徴の組合せの数は膨大であるため，有効な特徴の組合せを効率よく探索するためのルールマイニング手法を提案する．提案手法は，逐次的に特徴を探索しながら有効な特徴の組合せを発見することで，処理時間を削減する．

2. 関連研究

渡辺ら [1], [5] の研究では，イベント検出の一環として，ツイートに出現したスポット名から，そのスポットの緯度経度情報をツイートに結び付ける手法を提案している．位置情報サー

ビス Foursquare(注1) を利用し，スポット名と緯度経度情報の組合せからスポット情報データベースを構築する．次に，スポット情報データベースに登録してある一つのスポット名に対して複数の緯度経度情報が登録してある場合，地理的分散が大きければ，スポット情報データベースから除去する．そして，スポット情報データベースからスポット名の辞書を作成し，ツイートにスポット名が出現していた場合，そのスポット名の緯度経度情報をツイートに付与する．榊 [7] らの研究では，ツイートに対して，形態素解析を行った際の品詞情報やパターンマッチングから，スポット名を検出する．そして，Google Maps API(注2) を用いて，スポット名を緯度経度情報に変換して，ツイートに付与する．これらの研究では，ツイートに結び付けた緯度経度情報がツイート投稿時のユーザの位置情報かどうかの判定を行っていないため，ツイート投稿時のユーザの位置情報を利用するようなアプリケーションでは問題となる．本研究では，この問題に対応するため，付与した位置情報が投稿時のユーザの位置情報かどうかの判定を行う．

また，ユーザの居住地を推定する研究 [2], [3], [4] も行われている．これらの研究では，ユーザが投稿した一連のツイートや Twitter におけるユーザ間の関係を利用して，ユーザが普段住んでいる地域を推測する．ユーザの居住地に関する情報は，ツイートの投稿位置の推測に役立つと考えられるが，本研究ではツイート単体の特徴のみからどの程度正確にツイートの投稿位置を推測できるかを検証する．なお今後，ユーザの居住地を考慮した手法についても検討する予定である．

3. 先行研究とその課題

本章では筆者らの先行研究 [6] について説明した後，その課題について触れる．

3.1 先行研究の手法

先行研究では，教師あり機械学習手法である線形 SVM（Support Vector Machine）を用いて，ツイートに含まれるスポット名に対応する位置情報から実際にそのツイートが投稿されたかどうかを判定する手法を提案している．先行研究の手法の概要を図 1 に示す．まず，学習フェーズにおいて，位置情報付きツイートに対してジオコーディングを行う．ジオコーディングはツイートに含まれるスポット名を CRF（Conditional Random Field）により検出し，Foursquare から抽出したスポット名と位置情報をもとに，対応する位置情報を付与する．付与された位置情報と実際のツイートの位置情報との誤差距離から，付与された位置情報が投稿時のユーザの位置情報を表しているかを判定し，正解/不正解のラベル付けを機械的に行う．このときに，ツイートの文字数やスポット名のカテゴリ，ツイート中に出現する自立語など，様々な特徴をツイートの特徴量として，教師データを作成する．学習するツイートの特徴量は様々なものが考えられ，より多くの特徴を利用すればそれだけ精度向上が期待できる．なお，位置情報付きツイートから教師データを

(注1)：https://ja.Foursquare.com/
(注2)：https://developers.google.com/maps/

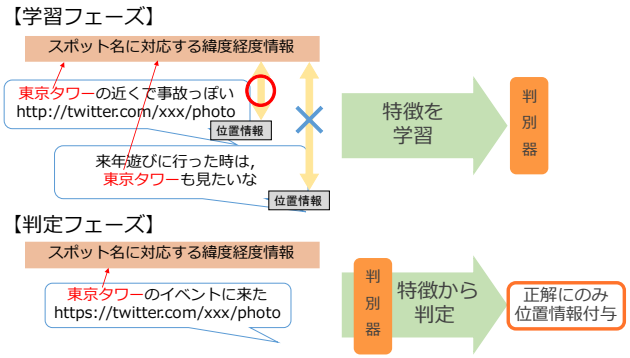

図 1　先行研究における手法の概要

作成すると，位置情報が付いていないツイートとは異なる特徴を学習し判定の精度が下がる恐れがある．これを緩和するため，位置情報付きツイートに特有の位置情報サービスやそれと連動したサービスから投稿されたツイートを除去する．

判定フェーズでは，学習フェーズで学習したツイートの特徴による判別器を利用して，実際にユーザがそのスポットからツイートを投稿したか否かを判定する．位置情報が付いていないツイートを入力とし，付与された位置情報が投稿時のユーザの位置情報であると判定された場合のみ，そのツイートに位置情報を付与する．判定に関して閾値を設定することにより，高い適合率で位置情報を付与できるため，これを位置情報付きツイートとみなすことで，位置情報付きツイートの総量を増やすことができる．

3.2 先行研究の課題

線形 SVM を用いた先行研究の手法では，ツイートの各特徴に対してそれぞれ重みを計算し，重みと特徴量の線形和として判定結果を出力する．したがって，重みを分析することにより個々の特徴がどの程度判定に寄与するかを見積ることはできる．一方，どのような特徴の組合せが判定に貢献するかについて分析できない．適合率を維持しつつ再現率をさらに向上させるためには，スポット名が指す位置情報からツイートが投稿されたか否かを判定するのに有効な特徴の組合せを詳細に分析する必要がある．この分析を行うためには，SVM などの機械学習手法では困難であるため，より直接的に分析する手法が必要となる．

4. 提 案 手 法

本研究では，先行研究の手法をもとに，ツイート中のスポット名が指す位置情報がツイートの投稿位置を表しているか否かを判定するのに有効な特徴の組合せを抽出する手法を提案する．提案手法は，大きく分けて教師データ作成と特徴の組合せ抽出の二つのフェーズに分割できる．4.1 節で教師データ作成，4.2, 4.3 節でそれぞれ適合率，エラー率の高い特徴の組合せの抽出について述べる．

4.1 教師データ作成

教師データの作成は，先行研究 [6] の手法を基に行う．まず，位置情報付きツイートに対しジオコーディングを行う．位置情報付きツイートに含まれる固有名詞をスポット名の候補とし，

あらかじめ作成しておいたスポット情報辞書を用いてスポット名に対応する緯度経度情報をツイートに付与する（実験で実際に使用したスポット情報辞書については 5.1.1 項で述べる）．ジオコーディングにより付与された緯度経度と位置情報付きツイートが持つ緯度経度との実空間上での距離をヒュベニの公式を用いて計算し，誤差距離が $1km$ 以内であれば正解，$10km$ 以上であれば不正解のラベルを付与する．なお，先行研究より，誤差距離が $1km$ 以上 $10km$ 以内であるツイートに関しては正解／不正解の判定が困難であるため，教師データとして使用しない．

ラベル付与の際，表 1 に示す各特徴に対して該当していれば 1，そうでなければ 0 の特徴量を持つ教師データを作成する．先行研究と異なる点として，表 1 においてツイートの文字数などのバイナリ値でないものについては，複数の閾値を用いてバイナリ値に変換した特徴を生成する．例えば，ツイートの文字数の場合，10 文字以下（以上）かどうか，といった特徴量に変換する．全ての特徴量をバイナリ値とすることにより，後述するルール抽出を可能とする．ツイートからスポット名が複数抽出された場合，それぞれのスポット名に対し位置情報を付与し，別々のツイートとみなして教師データを作成する．

位置情報付きツイートを用いることにより，大量の教師データを機械的に作成できるが，位置情報付きツイートは位置情報が付いていないツイートとは異なる特徴を持っている．その例として，位置情報連携サービスから投稿されたツイートやそれらに特有の定型文を持つツイートが挙げられる．このような位置情報付きツイートに特有なツイートをあらかじめ除去することで，教師データとする位置情報付きツイートと位置情報の付いていないツイートの特徴の差異を減らすことができる．なお，これを実現する方法としてはいくつか考えられるが，本研究では，5.1.2 項で示すように，自動的に投稿が行われにくいと考えられる Twitter クライアントを絞り込むことで実現する．

表 1　特 徴 一 覧

対象	特徴
ツイート	ツイートの文字数
	ツイートの文字数（URL，@部分，#を除去）
	ツイートの文字構成
	ツイートの文字構成（URL，@部分，#を除去）
	ツイートの品詞構成
	投稿時の時間帯
	投稿日の曜日
	投稿日が平日か休日か
	投稿クライアントの種類
	URL の出現回数
	メンションの出現回数
	#の出現回数
	RT かどうか
スポット名	スポット名自体
	スポット名の文字数
	スポット名の文字構成
	ツイート内で検出したスポット名の数
	スポット名のカテゴリ
	最初から何語句目か
	最後から何語句目か
語句	スポット名の 1 文字前の語句の品詞
	スポット名の 1 文字前の語句自体
	スポット名の 1 文字後の語句の品詞
	スポット名の 1 文字後の語句自体
	先頭の語句の品詞
	先頭の語句自体
	末尾の語句の品詞
	末尾の語句自体
	スポット名の係り受け先の語句自体
	スポット名の係り受け先の文字構成
	スポット名の係り受け先の品詞構成
	頻出語の語句自体

4.2　適合率の高い特徴の組合せの抽出

作成した教師データを直接分析することで，単独の特徴よりもツイート投稿位置の付与における適合率が高くなる複数の特徴の組合せを抽出する．ツイートの特徴は，表 1 をすべてバイナリ値で表現したとき，約 12,000 次元となった．この高次元の中から特徴の組合せを抽出しようとした場合，それらの組合せの数は膨大なものとなる．そこで，適合率が高くなる特徴の組合せを効率的に抽出する手法を提案する．以下では具体的なアプローチとアルゴリズムについて説明する．

4.2.1　アプローチ

提案手法のアプローチとして，相関ルール [8] をベースとした手法を考える．相関ルールとは，前提部となるある事象 X の下で，結論部となるある事象 Y が発生するという関係を表す．本研究においては，ある特徴の組合せを持つという事象の下，正解のフラグが付与されているという事象，つまりスポット名の位置情報と投稿時のユーザの位置情報とが一致するという事象が起こる確率が高くなるルールを，適合率が高くなる相関ルールとして抽出する．

教師データが大規模で且つ特徴の組合せの数が膨大であるため，厳密手法による探索は困難である．一般に大規模なデータでは近似手法 [9] を用いることで，効率よく探索を行える．本研究では，大規模かつ高次元な教師データから，判定に有効な特徴の組合せを発見する近似手法を提案する．提案手法は，適合率が高くなる特徴の組合せを効率的に抽出するため，適合率の高い特徴から順に特徴を一つずつ追加しながら組合せを探索する．ある特徴の組合せに別の特徴を追加する際には，追加する前より適合率が増加する場合にのみ追加する．これにより，特徴を一つずつ追加したときに適合率が単調増加するような特徴の組合せを網羅的に発見できる．なお，三種類以上の特徴を組み合わせてはじめて有効に機能するような組合せを発見できないが，線形 SVM を用いた判定がうまく機能していたことを考えると，逐次的に特徴を追加する手法でもある程度うまく機能すると考えられる．また，組合せの出現回数（サポート値）に閾値を設けることで，処理を高速化しつつ，ルールとして信頼できるもののみを抽出する．最終的に，指定した適合率以上となる特徴の組合せをルールとして得る．ルールに合致するツイートに対して位置情報を付与することで，高い適合率でツ

イート投稿時のユーザ位置推定を行える．

提案手法で用いる適合率，エラー率，およびサポート値について定義する．次元数 n の特徴 $f_1, \cdots, f_n$ に対して，特徴の組合せ（あるいは単体の特徴）C における適合率 $P(C)$，エラー率 $E(C)$，およびサポート値 $S_P(C)$ を以下の式により定義する．

$$P(C) = \frac{TP_C}{TP_C + FP_C}$$
$$E(C) = \frac{FP_C}{TP_C + FP_C} = 1 - P(C)$$
$$S_P(C) = TP_C$$

TP_C はすべての $f_i \in C$ についての特徴量の値が 1 かつ正解のフラグが付与されているツイートの数であり，FP_C はすべての $f_i \in C$ についての特徴量の値が 1 かつ不正解のフラグが付与されているツイートの数である．$P(C)$ は特徴の組合せ C 内のすべての特徴に対し該当する（特徴量の値が 1 の）ツイートのうち，実際に正解のフラグが付与されているツイートの割合を，$E(C)$ は不正解のフラグが付与されているツイートの割合を意味している．また，サポート値 $S_P(C)$ は TP_C として表される．

4.2.2 アルゴリズム

4.2.1 項のアプローチに基づくアルゴリズムの擬似コードを Algorithm 1 に示す．まず，n 個の特徴（$f_1, \cdots, f_n$）からサポート $S_P(f_k)$ が閾値 s 以上でかつ適合率 $P(f_k)$ が λ 以上の特徴 m 個（$F_1, \cdots, F_m$）にフィルタリングする．ここで，λ は単独の特徴での適合率の閾値であり，その特徴が含まれることで適合率が上がる可能性が高いもののみを候補とする目的で導入している．m 個の特徴を $P(F_k)$（$1 \leqq k \leqq m$）が高い順に並び替えた $List$ を作成し，F_k が $List$ において先頭からどの位置に格納されたかを参照できるように $List_{INDEX[f_k]}$ を構築する．これは，単独の適合率が最も高くなる特徴が $List_{INDEX[f_k]} = 0$ に，単独の適合率が最も低くなる特徴が $List_{INDEX[f_k]} = m - 1$ になるような関数である．組合せ C 内の特徴 $f_c \in C$ で最大の $N_{max} = List_{INDEX[f_c]}$ の値を求めることにより，C に追加しうる特徴の候補を List$[N_{max}]$ 以降の配列要素とすることができ，探索領域の削減を行える．各ループ内では，特徴の組合せ $C \in A$ について，特徴を一つだけ C に追加したときに，$P(C_{new}) > P(C) + \epsilon E(C)$ かつ $S_P(C) \geqq s$ を満たす場合のみ，特徴を追加した新たな組合せを配列 A の要素とする．なお，$\epsilon E(C)$ を加算して適合率を比較しているのは，エラー率がある程度改善された場合にのみ，適合率が増加したとみなすためである．これがない場合，特徴を追加することで適合率に実質的な変化がない場合でも，ごくわずかに適合率が上昇する場合が頻出するため，探索領域が大きくなる．各ループで A 内の全ての組合せを処理していき，C_{new} の適合率が閾値 θ を超える場合に C_{new} をルール集合 R に追加する．最終的に，R に含まれる特徴の組合せはすべて閾値 θ よりも高い適合率となる．

Algorithm 1 適合率の高いルールの抽出アルゴリズム

入力: サポートの閾値 s，単独の特徴での適合率の閾値 λ，ルールの適合率の閾値 θ

$[F_1, \cdots, F_m] \Leftarrow [f_1, \cdots, f_n]$ SATISFYING $S_P(f_k) \geqq s$ and $P(f_k) \geqq \lambda$
$List = [F_1, \cdots, F_m]$ SORTED BY $P(F_k)$
BUILD $List_{INDEX}[f_k]$ MEANS *Index of* F_k *in List*
$A \Leftarrow List$, $R \leftarrow []$
loop
  $Y \Leftarrow []$
  for C in A **do**
    $N_{max} = max(List_{INDEX[F_c]}) : F_c$ is any feature in C
    for $i = N_{max} + 1$ to $n - 1$ **do**
      $C_{new} \Leftarrow C \cup List[i]$
      if $P(C_{new}) > P(C) + \epsilon E(C)$ and $S_P(C_{new}) \geqq s$ **then**
        $Y \Leftarrow Y \cup C_{new}$
        if $P(C_{new}) > \theta$ **then**
          $R \leftarrow R \cup C_{new}$
  $A \Leftarrow Y$

出力: ルール集合 R

4.3 エラー率の高い特徴の組合せの抽出

4.2 節と同様に教師データを分析することで，適合率が高い特徴を含む組合せのうち，エラー率が高くなる特徴の組合せを抽出する．以下で具体的なアプローチとアルゴリズムについて説明する．

4.3.1 アプローチ

SVM により正解と判定されたツイートのうち，エラー率が高い特徴の組合せに該当するツイートをフィルタリングすることによって，同等の適合率を達成しつつ再現率を向上させることができると考えられる．SVM では個々の特徴量と重みの線形和による判別を行うため，単独の特徴では有効に機能するものの，他の特徴と組み合わさることによってエラー率が高くなる特徴の組合せを直接的に分析することができない．そこで，単独で適合率が高い特徴と，その他の特徴との組合せを網羅的に探索し，エラー率が高くなる特徴の組合せを抽出する手法を提案する．なお，単独での適合率が低い特徴同士の組合せを探索しないのは，そのような特徴のみを持つツイートは SVM によって正解と判定されず，フィルタリングする必要がないためである．

適合率が高い特徴を含む組合せのうち，エラー率が高くなる特徴の組合せを抽出するため，単独での適合率が閾値以上となる特徴と，その他の特徴との組合せを探索する．問題の性質上，エラー率が高い特徴の組合せは，適合率が高い特徴の組合せよりも圧倒的に数が多いため，本研究では，二種類の特徴の組合せ（特徴のペア）のみを網羅的に探索する．そして，指定したエラー率以上となる特徴のペアをルールとして抽出する．得られたルールに合致するツイートを，SVM により正解と判定されたツイートから除去する．

エラー率の高い特徴の組合せの抽出では，サポート値を以下により定義する．

$$S_E(C) = FP_C$$

これは，不正解とラベル付けされたツイートの数をサポート値とすることを意味している．また，適合率，エラー率は 4.2.1 項と同じ定義である．

4.3.2 アルゴリズム

4.3.1 項のアプローチに基づくアルゴリズムの擬似コードを Algorithm 2 に示す．まず，n 個の特徴（$f_1, \cdots, f_n$）から $P(f_k)$ が単独の特徴での適合率の閾値 λ 以上の特徴 m 個（$F_1, \cdots, F_m$）を抽出する．λ によって，単独の特徴での適合率が高いもののみに絞り込む．m 個の特徴と，自身を除く $n-1$ 個の特徴との組合せを網羅的に探索し，エラー率が閾値 θ を超える場合にのみ，組合せ C として保持する．各ループ内では，特徴 F_k に特徴 f_k（$F_k \neq f_k$）を組み合わせた C に対して，$E(C) > \theta$ かつ $S_E(C) \geqq s$ を満たす場合にのみ，C を R に追加する．最終的に，R に含まれる特徴の組合せはすべて閾値 θ よりも高いエラー率となる．

Algorithm 2 エラー率の高いルールの抽出アルゴリズム

入力: サポートの閾値 s，単独の特徴での適合率の閾値 λ，
ルールの適合率の閾値 θ

1: $[F_1, \cdots, F_m] \Leftarrow [f_1, \cdots, f_n]$ SATISFYING $S_E(f_k) \geqq s$ and $P(f_k) \geqq \lambda$
2: $R \leftarrow []$
3: **for** F_k in $[F_1, \cdots, F_m]$ **do**
4: **for** f_k in $[f_1, \cdots, f_n]$ **do**
5: **if** $F_k \neq f_k$ **then**
6: $C = F_k \cup f_k$
7: **if** $E(C) > \theta$ and $S_E(C) \geqq s$ **then**
8: $R \leftarrow R \cup C$

出力: ルール集合 R

5. 評価実験

特徴の組合せに関するルールを用いた判定結果と SVM による判定結果との比較評価を行い，抽出したルールの有効性を検証した．また，実際に抽出したルールについて分析を行った．

5.1 データセット

5.1.1 スポット情報辞書

Yahoo! Open Local Platform（YOLP）(注3) が提供する Yahoo! ローカルサーチ API から取得した POI（Point Of Interest）情報からスポット名と緯度経度情報を抽出し，Yahoo!スポット名としてスポット情報辞書を作成した．このスポット情報辞書は，チェーン店名とその店舗名をスペース区切りで繋げたスポット名が多く登録されていた．この形式のスポット名は，スポット名をスペースで分割した後半部分の「〜店」の「〜」部分は，スポット名となっている場合が多かったため，「店」の表記の部分を削除したスポット名を分割スポット名としてスポット情報辞書に登録した．また，国土交通省国土地理院の位置参照情報ダウンロードサービス (注4) から，範囲が狭い街区レベルの地名と代表点の緯度経度情報を取得し，街区スポット名としてスポット情報辞書に追加した．また，Yahoo! ローカルサーチ API では駅名を網羅的に取得することができなかったため，駅.jp(注5) が提供する API から各路線毎の駅名と緯度経度情報を取得し，駅スポット名としてスポット情報辞書に追加した．作成したスポット情報辞書から，各スポット名に対応する緯度経度情報とそのカテゴリを取得できる．

5.1.2 位置情報付きツイート

Twitter Streaming API(注6) を用いて，2015 年 10 月 1 日から 2016 年 1 月 31 日に日本を含む矩形領域内で投稿された位置情報付きの日本語のツイートを取得した．bot による自動投稿や携帯端末以外からの投稿，位置情報付きツイートに特有の位置情報サービスやそれと連動したサービスから投稿されたツイートを除去するため，Twitter クライアントによる絞り込みを行った．Twitter for Android，Twitter for iPhone，Janetter for Twitter on iOS など計 21 件のクライアントによるツイートのみを対象とし，それ以外のツイートをノイズとして除去した．2015 年 10 月 1 日から 2016 年 1 月 26 日に投稿されたツイート 7,949,313 件から，ジオコーディングにより緯度経度情報を付与し，誤差距離から正解/不正解のラベル付けができたツイート 275,641 件 (注7) を教師データとして用いた．教師データの内，実際の位置情報との誤差が 1km 以下で正解のラベルを付与した回数は 17,656 回，10km 以上で不正解のラベルを付与した回数は 257,985 回であった．正解のラベル付けを行ったツイートは位置情報付きツイートの約 0.22%に当たる．また，テストデータとして 2016 年 1 月 27 日から 2016 年 1 月 31 日に投稿されたツイート 380,530 件のうち，ジオコーディングにより緯度経度情報を付与し，正解/不正解のラベル付けができたツイート 10,069 件を用いた．テストデータの内，正解のラベルを付与した回数は 596 回，不正解のラベルを付与した回数は 9,473 回であった．

5.2 適合率が高い特徴の組合せの評価

5.2.1 提案手法により抽出された特徴の組合せ

提案手法により教師データに対して適合率を高く判定できる特徴の組合せを抽出した後，テストデータにおいて，抽出した組合せが該当するツイートに対して位置情報を付与したときの適合率の評価を行った．単独の特徴での適合率の閾値 λ を 0.1，サポート値の閾値 s を 100 または 10 として抽出した組合せにおいて，指定した閾値 θ 以上の適合率となる特徴の組合せに対する評価結果を図 2，3 に示す．なお，各特徴の組合せに該当するツイートの重複を含む場合，含まない場合についてそれぞれ適合率を求めた．

サポート値が 100 の場合と 10 の場合での適合率を比較すると，サポート値が 100 の場合の方が総じて適合率が上回っている．これは，サポート値が 100 の場合の方がより信頼性の高いルールを抽出できたことを意味している．また，重複を含む場合，各ルールの教師データにおける適合率はすべて指定した閾値以上となるため，サポート値が十分に大きく，抽出したルールとしての信頼性が高ければ，テストデータにおける適合率はその閾値より大きくなる可能性が高いといえる．

(注3)：http://developer.yahoo.co.jp/webapi/map/
(注4)：http://nlftp.mlit.go.jp/isj/index.html
(注5)：http://www.ekidata.jp
(注6)：https://dev.twitter.com/streaming/overview
(注7)：ツイート内に複数のスポット名が出現している場合，それぞれのスポット名について 1 ツイートとカウントしている．

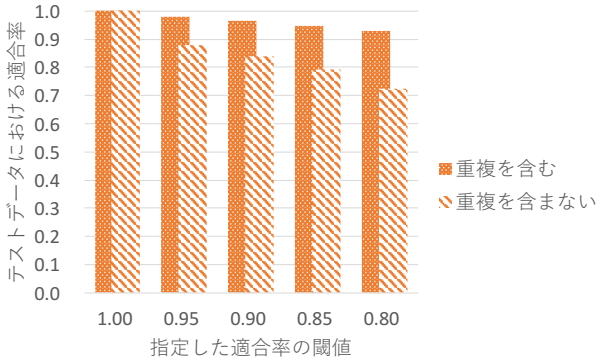

図 2　サポート値:100 における適合率の評価

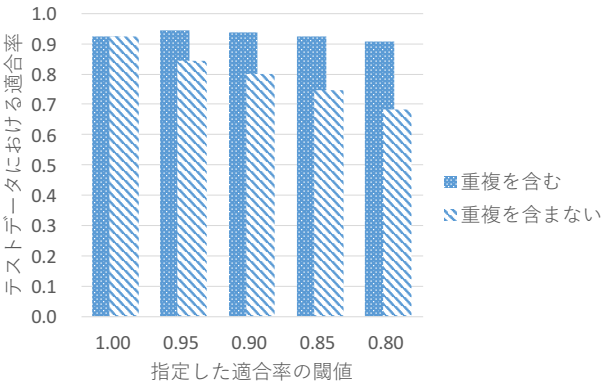

図 3　サポート値:10 における適合率の評価

一方，ツイートの重複を含まない場合，すなわち，ルール集合全体として位置情報を付与したツイートについて適合率を算出した場合，閾値の適合率よりも下がる傾向にあることがわかった．これは，位置情報を正しく付与できたツイートはルールが重なっていることが多かったためである．このことから，ルール集合全体としての適合率を指定した閾値以上とするためには，教師データを分割し，ルールについて交差検証を行うなどの工夫が必要である．

5.2.2　提案手法と線形 SVM を用いる手法の比較

次に，テストデータにおいて線形 SVM を用いる手法 [6] と比較したときの適合率および再現率を評価した．線形 SVM の評価には，liblinear(注8) [10] を用いた．ツイートの特徴量はスパース性が高いため，正則化手法は L1 正則化を使用した．また，liblinear の機能であるグリッドサーチを用いて，最適なコスト c を 0.0625 とした．図 4 に評価結果を示す．提案手法の各点は，指定した適合率の閾値を 1.0 から 0.01 ずつ変化させたときのテストデータに対する適合率，再現率をプロットしたものである．また，比較手法（線形 SVM）の各点は，正解判定における重みのスコアの閾値を，0.025 ずつ変化させたときのテストデータに対する適合率，再現率をプロットしたものである．

適合率が約 0.82 以上のとき，提案手法はいずれも比較手法と比べて再現率が向上している．このことから，高い適合率で位置情報を付与しようとした場合には，提案手法により抽出し

(注8)：https://www.csie.ntu.edu.tw/ cjlin/liblinear/

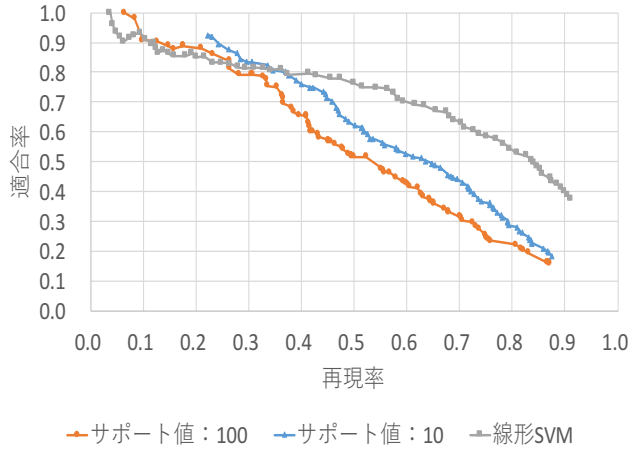

図 4　適合率が高い特徴の組合せの抽出の評価結果

たルール（特徴の組合せ）を用いることで，より多くのツイートに対して位置情報を付与できる．特に適合率 1.0 のときは，先行研究と比較しサポート値の閾値を 100 と高く設定することで，線形 SVM を用いた比較手法に対し 2 倍近い再現率を達成できた．また，適合率が 0.9 程度の場合でも，線形 SVM では再現率が約 0.11 であったのに対し，サポート値が 10 の場合に約 0.22 の再現率を達成できている．

一方で，適合率が 0.8 以下の低い値のとき，提案手法における再現率は線形 SVM と比べて著しく低下していることがわかる．これは，高い適合率を達成したい場合にのみ特徴の組合せが有効に機能するためであると考えられる．適合率と再現率のバランスを重視したい場合は，線形 SVM を用いて個々の特徴の重みを最適化させた判別器を用いるほうが有効であるといえる．サポートの値を 10 として抽出した適合率 0.9 以上の特徴の組合せを持つツイート集合と，線形 SVM によって 0.9 以上の適合率で正解と判定したツイート集合とを比較する．テストデータの正解ツイート 596 件，不正解ツイート 9,473 件に対し，提案手法では 153 件のツイートを正解と判定し，139 件のツイートが実際に正解であった（139/153 = 0.908）．また，線形 SVM では 80 件のツイートを正解と判定し，72 件のツイートが実際に正解であった（72/80 = 0.900）．したがって，提案手法と線形 SVM によって得られたそれぞれのツイートの和集合は 192 件となり，そのうち実際に正解であったツイート数は 171 件（171/192=0.891）であった．提案手法と線形 SVM とを組み合わせることにより，適合率を 0.9 前後に維持したまま，位置情報を付与できるツイートの数を増やすことができる．

抽出した具体的な特徴の組合せについて考察する．表 2 に適合率が高い特徴の組合せと該当するツイートの例を示す．提案手法により抽出した特徴の組合せで教師データにおける適合率 0.9 以上かつサポート値 10 以上の例として，(単語の数が 3 つ，スポット名が 1 文字目，URL の数が 1 つ以上)，(スポット名が ‘羽田空港’，スポット名が分割スポット名に合致，URL の数が 1 つ以上，スポット名の 1 単語後が ‘で’)，(スポット名が ‘赤羽’，スポット名の 3 単語後が ‘より’，自立語 ‘店’ を含む）などの組合せを抽出した．抽出した組合せの傾向から，個

表 2　適合率が高い特徴の組合せと該当するツイートの例

ツイート	特徴の組合せ
大阪城公園到着！ https://t.co/Jjgp0vpi5m	(単語の数 3 つ，スポット名が 1 文字目，URL の数 1 つ以上)
赤羽 2 号店より本日日替りは US 産ポークソテー 930 円！ スペシャルプライスです。	(スポット名' 赤羽'，スポット名の 3 語句後 ' より '， 自立語 ' 店 ' を含む)
中央公園なう!つくば駅 A2 出口すぐそば, 交番裏の白いテントがもっくんカフェです	(駅スポット名，自立語' なう'，URL の数 1 つ以上)
まじで、いまやっと宇都宮駅着いた 模試が 6 時までっていう鬼畜なことしてきた ほんと病みそうだったテスト何にも わからなかったぜ もううちは受けなくていいや笑笑	(駅スポット名，スポット名の 1 語句後' 着い')
羽田空港でアンケート受けたらいただいた。 本気で買おうか迷っていたし 記念になるからいいねー！ https://t.co/x0lxc92jrK	(スポット名' 羽田空港'，分割スポット名，URL の数 1 つ以上, スポット名の 1 語句後' で')
越後中里駅目の前の湯沢中里スキー場で無料休憩室に使われている旧客たち。 全部で 17 両ぐらい置いてあります https://t.co/6k626LMdU2	(駅スポット名，スポット名が漢字 5 文字，URL の数が 1 つ以上 スポット名が 1 文字目)

人の一般ユーザがスポット名に関連した投稿をするものに加えて,「赤羽 2 号店より 本日日替りは US 産ポークソテー 930 円！スペシャルプライスです。」のように飲食店などのアカウントから投稿されるツイートに当てはまるルールを確認できた．また，ルールによって位置情報を付与できたツイートについて線形 SVM と比較すると，文字数の多いツイートについて正解と判定できている割合が多かった．線形 SVM で学習した特徴の重みを分析すると，ツイートの文字数が少ない場合には重みが大きく，文字数が多い場合には重みが小さくなっていた．このことから，SVM では文字数の大小で重みの差が大きく，文字数が多いツイートは重みの影響により正解と判定されにくくなる．提案手法では，抽出したルールに含まれる特徴のみを考慮するため，組合せに入っていない特徴についてはその影響を受けない．したがって，文字数が多いツイートに関しても，適合率が高い特徴の組合せを含むツイートは正解と判定できたと考えられる．

5.3　エラー率の高い特徴の組合せの抽出

正解データに対してエラー率が高くなる特徴の組合せについても評価を行った．サポート値が 10，単独の特徴での適合率が 0.3 以上の特徴と他の特徴との組合せを網羅的に探索し，特徴を組み合わせた時のエラー率が 0.9 以上となる特徴のペアを 627 件抽出した．テストデータから，抽出した特徴の組合せに該当するツイートを除去すると，正解ツイートの 19 件，不正解ツイートの 185 件がそれぞれ除去された．図 5 に，ツイートを除去せずに SVM による判定を行ったもの（先行研究による手法）と，除去した後に SVM による判定を行ったものについての評価結果を示す．提案手法によりエラー率が高くなる特徴の組合せを除去したものの，適合率の向上は見られなかった．特徴を 2 つまでしか考慮しない現状の探索手法では，エラー率の増加に寄与する特徴の組合せを効果的に探索できていない可能性が高い．

6.　ま　と　め

本稿では，位置情報付きツイートの総量を増やすことを目的とし，できる限り正確に位置情報を推測できるツイートのみをうまく判別してツイートの投稿位置を付与する手法を提案した．提案手法では，位置情報付きツイートから機械的に教師データを作成し，位置推定に有効なツイートの特徴の組合せ（ルール）を，膨大な数の組合せから効率よく探索する．提案手法により，適合率の高いルールに該当するツイートのみを判別することで，高い適合率を維持しつつ，線形 SVM を用いた比較手法よりも再現率を向上させることができた．また，線形 SVM と提案手法を組み合わせて使用することにより，より多くのツイートに位置情報を付与できた．

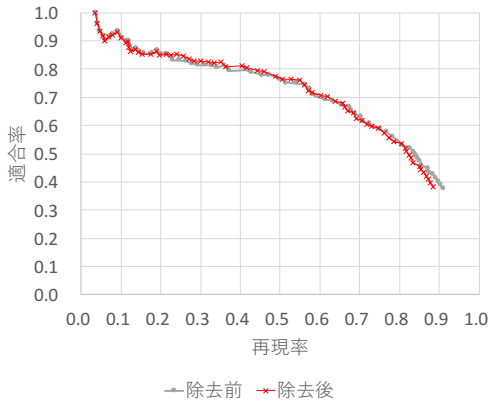

図 5　エラー率の高い特徴の組合せ抽出の評価

今後の研究課題として，エラー率の高い特徴の組合せの抽出の改善が挙げられる．本研究では，組み合わせる特徴の数を 2 つに絞って探索を行ったが，エラー率を削減できるような有効なルールは抽出されなかった．組み合わせる特徴の数を増やし，より網羅的に探索を行うことで，有効なルールが抽出できる可能性がある．これを行うためには，エラー率の高い特徴の組合せを効率的に探索する手法が必要となる．また別の課題として，本研究では位置情報付きツイートに対してのみ評価を行ったが，今後，位置情報が付いていないツイートにおいて，抽出したルールがどの程度有効に機能するかを評価する予定である．

謝　　辞

本研究の一部は，文部科学省科学研究費補助金・基盤研究(A)(26240013)，JST 国際科学技術共同研究推進事業（戦略的国際共同研究プログラム），文部科学省国家課題対応型研究開発推進事業－次世代 IT 基盤構築のための研究開発－「社会システム・サービスの最適化のための IT 統合システムの構築」の研究助成によるものである．ここに記して謝意を表す．

参 考 文 献

[1] 渡辺一史，大知正直，岡部誠，尾内理紀夫，“Twitter を用いた実世界ローカルイベント検出”，楽天研究開発シンポジウム，2011.

[2] Z. Cheng, J. Caverlee and K. Lee, “You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users”, Proc. of CIKM, pp.759-768, 2010.

[3] H.-W. Chang, D. Lee, M. Eltaher and J. Lee, “@Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage”, Proc. of ASONAM, pp.111-118, 2012

[4] C. Davis, G.-L. Pappa, D.-R.-R. de Oliveira, and F. de L. Arcanjo, “Inferring the Location of Twitter Messages Based on User Relationships”, T. GIS, Vol.15, No.6, pp.735-751, 2011.

[5] K. Watanabe, M. Ochi, M. Okabe and R. Onai, “Jasmine: A Real-time Local-event Detection System Based on Geolocation Information Propagated to Microblogs”, Proc. of CIKM, pp.2541-2544, 2011.

[6] 杉谷卓哉，白川真澄，原隆浩，西尾章治郎，“教師あり機械学習を用いたツイート投稿時のユーザ位置推定手法”，情報処理学会研究報告，Vol.2013，No.26，pp.1-8，2013

[7] 榊剛史，松尾豊，“ソーシャルメディアからの人物目撃情報抽出システムの試作”，人工知能学会全国大会 2011 論文集，Vol.25，pp.1-4，2011.

[8] R. Agrawal, T. Imielinski, A. Swami, “Mining Association Rules Between Sets of Items in Large Databases”, Proc. of SIGMOD, 1993

[9] F. Zhu, X. Yan, J. Han, P.-S. Yu, H. Cheng, “Mining Colossal Frequent Patterns by Core Pattern Fusion”, Proc. of ICDE, pp.706-715, 2007

[10] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, Chih-Jen Lin, “LIBLINEAR: A Library for Large Linear Classification, Journal of Machine Learning Research”, Vol. 9, pp.1871-1874, 2008.