機械学習による中国語助詞の用法解析

宋 東旭 (東京農工大学 工学府) *
浅原 正幸 (国立国語研究所 コーパス開発センター)
古宮 嘉那子 (東京農工大学 工学研究院)
小谷 善行 (東京農工大学 工学研究院)

Comparison of Resampling Strategies for Chinese Auxiliary Word Classification

Dongxu Song (Graduate School of Engineering, Tokyo University of Agriculture and Technology)
Masayuki Asahara (Center for Corpus Development, NINJAL)
Kanako Komiya (Institution of Engineering, Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani (Institution of Engineering, Tokyo University of Agriculture and Technology)

1. はじめに

現代中国語文法の助詞には「語気助詞」、「時態助詞」と「結構助詞」の三種類ある。結構助詞は、日本語の接続助詞にあたり、「的」、「地」、「得」などがこの分類に入る。いずれも/de/と発音され計算機上の入力方法が同じであるため、誤用・混用が多い。

本研究では結構助詞の/de/が出現する位置に「的」、「地」、「得」のいずれの語を選択すべきかを判定する機械学習方法を提案する。対象となる結構助詞の前後二形態素の文脈情報を用い、サポートベクトルマシン (Support Vector Machines; 以下 SVM) (Vapnik (1995)) と多クラスロジスティック回帰（最大エントロピー法）を利用することで入れるべき語を選択する。いずれの手法も分布の偏りに弱く、一方のデータが多い場合にはそちらに引きずられる傾向にある。そこで訓練事例として与えるクラスラベル間の分布を調整することで少ない事例に対応できるかどうかを試みた。SVM は訓練事例が分布する素性空間上の分離超平面を探索する識別学習器であり、ロジスティック回帰は訓練事例がラベルに条件づけられて二項分布で生成されたものとする統計学習器である。それぞれでサンプル数の増減が識別器にどのような影響を与えるのかを検証した結果を報告する。

2. 背景および関連研究

2.1 「的」「地」「得」

「的」はヒト・モノ・コトがどのような態度・状態・程度にある（静態）かを表現するときに利用し、一般に被修飾名詞に前置する。「地」はヒト・モノ・コトがどのような態度・状態・程度で動く（動作）かを表現するときに利用し、一般に被修飾動詞に前置し、修飾形容詞に後置する。「得」はヒト・モノ・コトがどのような態度・状態・程度にする（到達・結果）かを表現

* songdongxu123@gmail.com

するときに利用し、一般に修飾動詞に後置する。

この三種類の結構助詞のうちテキストの出現の観点では「的」が大勢を占めており分布に偏りがある。特に「地」は「的」と誤用され,「得」は「地」,「的」と誤用される (蒋 (2005))。以下にそれぞれの誤用の例を示す。

飞快**的**奔跑	飞快**地**奔跑の誤り	(速く走る)
跑**地**很快	跑**得**很快の誤り	(走るのがとても速い)

現在では中国語学者にも “全て「的」を使用するようにする” という意見もある (王 (2009))。

2.2 機械学習による表現選択

言語処理の分野において機械学習技術を用いた表現選択手法は数多く行われている。古宮ほか (2008) は決定木学習により適切な敬語を選択する規則の獲得手法を提案した。この研究における実験では、判定対象である普通語：尊敬語：謙譲語の比率が 60%:27%:13% もしくは普通語：丁寧語の比率が 49%:51% と比較的均衡が取れたデータであった。竇ほか (2012) はサポートベクトルマシンを用いて中国語のネット語を判定システムを構築した。この研究における評価実験ではネット語 5000 文に対し書き言葉 2000 文で実験しており、極端に分布が偏ったものではない。一方、本研究が対象とする「的」「地」「得」は 95% が「的」である偏った分布であり、これらの先行研究と異なった方策が必要であると考える。次節では分布が偏りがあるデータに対する識別学習の先行研究について示す。

2.3 分布に偏りがあるデータに対する識別学習

分布に偏りのあるデータを Unbalanced Data もしくは Imbalanced Data と呼ぶ。SVM をこのようなデータに適応するにあたり、さまざまな手法が提案されている。その中でデータ集合の均衡を持たせる手法に着目する。Kubat and Matwin (1997) はサンプル数が少ないクラスのデータ量を保持したまま、サンプル数が多いクラスのデータ量を削減することによりデータのバランスを取る方法 (UnderSampling 法) を提案している。UnderSampling 法では、サンプル数が多いクラスのデータを損失してしまうという問題点がある。この問題点に対して少ないクラスのデータ量を重複化させたり、素性空間上で線形補間したりして増やす方法 (OverSampling 法) が提案されている (Japkowicz and Stephen (2002);Chawla et al. (2000))。しかし、SVM の分離平面探索手法の性質から、OverSampling 法ではあまりよい性能が得られず、一般に UnderSampling 法の方がよいことが知られている。そこで Wang and Japkowicz (2009) は UnderSampling 法と OverSampling 法の分布の粒度を変えた SVM を複数作成したものを Boosting する Boosting SVM を提案した。

3. 比較する手法

SVM は二値分類器であり、複数種類のクラスラベルに分類するには one-against-others 法と one-against-one (pairwise) 法などがある (Hsu and Lin (2002)) が、利用する SVM のパッケージ LibSVM は後者を採用している。多クラスロジスティック回帰は正則化手法として L2 正則化と L1 正則化の二つの手法があり、利用する SVM のパッケージ LibLinear はこの両者に対

応している。

本研究では分布が極端に偏った “DE” 相当語の識別について、SVM で提案されているサンプル数を変更する手法の比較を試みる。SVM、L2 正則化多クラスロジスティック回帰、L1 正則化多クラスロジスティック回帰の三つの学習手法について、以下の五つの設定で実験を行い比較する。

- 通常のサンプル分割 (Normal) : 訓練データとテストデータをそのままの分割で訓練する。
- UnderSampling : 訓練データ中最も多いクラスラベル「的」について、20% に削減し、それ以外のクラスラベルについては元のサンプル数のまま訓練する。
- OverSampling : 訓練データ中少ないクラスラベル「得」「地」について、500% に増加し、それ以外のクラスラベルについては元のサンプル数のまま訓練する。
- UnderSampling の多数決 (US Vote) : 訓練データ中最も多いクラスラベル「的」について、20% に削減し、それ以外のクラスラベルについては元のサンプル数のまま訓練した設定で五つ (20% × 5 = 100%) 作成し、その結果の多数決を取る。
- Normal, OverSampling, US Vote の多数決 (Vote) : 上の分割による Normal, OverSampling, US Vote の三つの実験結果の多数決を取る。

4. 評価実験

4.1 実験設定

実験データとして「人民日報タグ付きコーパス (PFR コーパス)」の 1998 年 1 月の記事を用いる。対象となる表現として品詞タグが “u” である「的」「地」「得」を対象とする。表 1 に実験データの分布を示す。

表 1　実験データの分布

ラベル	件数	(割合)
「的」	54487	(95.0%)
「地」	2156	(3.7%)
「得」	661	(1.1%)
合計	57304	

素性として、位置別の前後 2 形態素の単語、品詞情報を固定して用いる。実験において訓練データとテストデータの分割は記事の日付単位の 5 分割交差検定による。

実験設定として五つのサンプル分割を比較する。一つ目は通常のサンプルに基づく識別学習である。二つ目はサンプル数が多い「的」について 500% の (UnderSampling) を行ったものである。「的」のサンプルを五つにデータ分割し、「地」「得」はそのままのものであるものと組み合わせたデータ集合を五つつくり、その結果の平均を取った。三つ目はサンプル数が少ない「地」と「得」について 500% の (OverSampling) を行ったものである。尚、特にサンプル間の線形補間を行ってサンプル数を増やすものではなく、単純な重複化 (duplication) によるものである。四つ目は UnderSampling で評価した五つのデータ集合の結果を多数決を取ったもの

(US Vote) である。五つ目は Normal, OverSampling, US Vote の 3 つのモデルの出力の多数決を取ったもの (Vote) である。

評価は全体の正解率とラベルごとの再現率、精度、F 値による。全体の正解率はラベルが適合したもの/全体の事例数である。再現率はそのラベルで正解した件数/訓練データ中の当該ラベル件数、精度はそのラベルで正解した件数/システムの当該ラベル出力件数、F 値は再現率と精度の調和平均である。

4.2 SVM

SVM のパッケージとして LibSVM を用いる。カーネルは線形カーネルを利用し、その他の設定はデフォールトの設定を用いる。表 2 に結果を示す。

表 2　実験結果 SVM (5 分割交差検定、マイクロ平均)

Normal	正解率 **99.17%**		
	再現率	精度	F 値
「的」	**99.70**	99.48	**99.59**
「地」	92.99	**93.86**	**93.42**
「得」	75.18	**88.59**	**81.34**

UnderSampling	正解率 98.62%		
	再現率	精度	F 値
「的」	98.93	99.69	99.31
「地」	96.09	83.29	89.23
「得」	81.55	72.01	76.49

OverSampling	正解率 99.13%		
	再現率	精度	F 値
「的」	99.67	99.47	99.57
「地」	93.04	93.34	93.19
「得」	74.58	88.03	80.75

US Vote	正解率 98.63%		
	再現率	精度	F 値
「的」	98.90	**99.74**	99.32
「地」	**96.56**	83.18	89.37
「得」	**83.35**	71.09	76.74

Vote	正解率 99.14%		
	再現率	精度	F 値
「的」	99.67	99.49	99.58
「地」	93.13	93.30	93.22
「得」	75.49	87.85	81.20

結果について考察する。まず全体の正解率については通常の SVM が最も性能が高い。OverSampling については、先行研究で言及されているように出力が通常の SVM から少し悪くなった。これは SVM がサンプルからの距離に基づいて分離超平面を設定する学習器であり、同じサンプルを増やすこと自体に本質的には意味がないからだと考える。今後、サンプル数が少ない事例の素性空間上の線形補間のような OverSampling 手法を検討する必要があるだろう。UnderSampling と US Vote は、サンプル数が少ない「地」「得」の再現率を高める一方、精度が悪くなり、結果として各ラベルの F 値および正解率が悪くなった。

4.3 L2 正則化ロジスティック回帰

ロジスティック回帰のパッケージとして LibLinear を用いる。L2 正則化オプション (-s 0) 以外はデフォールトの設定を用いた。表 3 に結果を示す。

結果について考察する。まず、全体の正解率については OverSampling と Vote が最もよかった。L2 正則化ロジスティック回帰は事前分布と事後分布がともに正規分布であることを仮定した事後確率最大化 (MAP) 推定を行う。未知データの対数尤度の期待値を訓練データの対数尤度で近似するが、UnderSampling のような訓練データの削減は元の分布でサンプル数が多い事例を少なくするバイアスをかけるだけでなく、単純に訓練サンプル数を削減するため性能が悪くなると考える。OverSampling は訓練サンプルを増やすために元のバイアスどおりの結果

表 3　実験結果 L2 正則化ロジスティック回帰 (5 分割交差検定、マイクロ平均)

Normal	正解率 99.01%		
	再現率	精度	F 値
「的」	**99.71**	99.32	99.51
「地」	92.25	92.68	92.46
「得」	63.08	91.04	74.53

UnderSampling	正解率 78.83%		
	再現率	精度	F 値
「的」	79.06	**99.71**	88.19
「地」	77.07	81.78	79.35
「得」	64.99	70.94	67.83

OverSampling	正解率 **99.03%**		
	再現率	精度	F 値
「的」	99.43	99.62	**99.53**
「地」	95.50	88.82	92.04
「得」	77.15	84.29	80.56

US Vote	正解率 98.45%		
	再現率	精度	F 値
「的」	98.77	99.69	99.23
「地」	**96.42**	80.11	87.51
「得」	**78.21**	71.31	74.60

Vote	正解率 **99.03%**		
	再現率	精度	F 値
「的」	99.44	99.62	**99.53**
「地」	93.13	**93.30**	**93.22**
「得」	75.49	**87.85**	**81.20**

が得られている。US Vote はこの訓練サンプル数を削減することによる欠点を多数決に補うため、サンプル数の少ない「地」「得」の再現率をあげるという元のバイアスどおりの結果が得られているが、精度が低い傾向にある。特に Vote についてはサンプル数の多い「的」の F 値を保ったまま、サンプル数の少ない「地」「得」の再現率にバイアスをかけたうえで、F 値/精度がよい傾向がみられる。

しかし、残念ながら L2 正則化そのものの結果は全体的に SVM に劣っており、サンプルの変化をしない SVM の方が性能がよいという結果になった。

4.4　L1 正則化ロジスティック回帰

前節の実験と同様にロジスティック回帰のパッケージとして LibLinear を用いる。L2 正則化オプション (-s 6) 以外はデフォールトの設定を用いた。表 4 に結果を示す。

表 4　実験結果 L1 正則化ロジスティック回帰 (5 分割交差検定、マイクロ平均)

Normal	正解率 98.97%		
	再現率	精度	F 値
「的」	99.66	99.34	99.50
「地」	91.88	**91.75**	91.81
「得」	65.05	**89.21**	75.24

UnderSampling	正解率 78.80%		
	再現率	精度	F 値
「的」	**99.71**	79.04	88.18
「地」	81.53	77.00	79.20
「得」	65.03	69.66	67.27

OverSampling	正解率 **99.00%**		
	再現率	精度	F 値
「的」	99.44	99.58	**99.51**
「地」	94.76	89.05	**91.82**
「得」	76.25	83.86	**79.87**

US Vote	正解率 98.29%		
	再現率	精度	F 値
「的」	98.61	**99.69**	99.15
「地」	**96.19**	79.13	86.83
「得」	**78.21**	66.11	71.66

Vote	正解率 98.98%		
	再現率	精度	F 値
「的」	99.46	99.56	**99.51**
「地」	94.48	89.18	91.75
「得」	74.43	84.54	79.16

結果について考察する。L1 正則化ロジスティック回帰は事前分布が Laplace 分布、事後分布が正規分布であることを仮定した事後確率最大化 (MAP) 推定を行う。学習時に不要な説明変数 (素性) に対する重みが 0 に縮退することができる。高次元の疎な素性空間に対して、説明変数を減らす場合には有効であるが、本タスクにおいては全体的に性能が悪くなった。これは、

前後 2 単語の素性を用いており素性空間が疎であるとはいえ、1 事例あたりの発火する素性は 8 つに限定されており、重みが 0 に縮退することで不定となる事例が多々あるからだと考える。L2 正則化と同様に US Vote がサンプル数の少ない「地」「得」の再現率をあげるバイアスがかけられている一方、精度が極端に悪くなる傾向にあり、Vote よりも単純な OverSampling がもっともよい正解率/F 値が得られることがわかった。

5. おわりに

本研究ではラベルの分布が偏った中国語助詞分類タスクにおいて、サンプル数を変えることによる機械学習器ごとと識別性能の変化を評価した。サンプル数が少ないラベルの再現率を上げるもっとも有効な方法として、サンプル数を減らした弱学習器の多数決を取る手法 (US Vote) がどの機械学習器にとっても有効であることがわかった。一方、全体の性能の観点からみると、SVM で通常のサンプル分割で学習するものが最もよかった。これは新聞記事という言語資源であっても、本来「的」を用いるべきではない部分に、「地」「得」を用いる傾向があるからではないかと考える。今後、識別学習の境界値情報をもつ Support Vector となった事例を検証しながら、元データの誤用例について言語学的な分析を進めていきたいと考える。

参考文献

Chawla, N., K. Bowyer, L. Hall, and W. P. Kegelmeyer (2000). “Smote: synthetic minority oversampling technique.” *International Conference on Knowledge Based Computer Systems*.

Hsu, C.-W., and C.-J. Lin (2002). “A comparison of methods for multi-class support vector machines.” *IEEE Transactions on Neural Networks*, pp. 415–425.

Japkowicz, N., and S. Stephen (2002). “The class imbalance problem: A systematic stydy.” *Intelligent Data Analysis*, 6:5.

Kubat, M., and S. Matwin (1997). “Addressing the curse of imbalanced training sets: One-sided selection.” *Proceedings of the 14th International Conference on Machine Learning*.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*.: Springer.

Wang, B. X., and N. Japkowicz (2009). “Boosting support vector machines for imbalanced data sets.” *Knowledge and Information Systems*.

王海峰 (2009).「区别使用“的，地，得”」 編集之友 2009 年 11 期.

古宮嘉那子・但馬康宏・小谷善行 (2008).「決定木を用いた敬語の選択ルールの獲得」 情報処理学会論文誌, 49:7, pp. 2679–2691.

蒋紹愚 (2005).『近代漢語研究概要』.

寶梓瑜・古宮嘉那子・小谷善行 (2012).「コーパスを用いた中国語ネット語の判定システム」第一回コーパス日本語学ワークショップ予稿集, pp. 161–166.

関連 URL

- LibSVM `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`
- LibLinear `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`