距離学習に基づく語義識別の性能分析

佐々木 稔　　新納 浩幸
茨城大学 工学部 情報工学科
{msasaki, shinnou}@mx.ibaraki.ac.jp

1　はじめに

ある単語が含まれる用例文集合に対して、語義別に用例文を分類することは本格的な意味解析を行う上で、非常に有用なデータセットの構築への可能性が広がる。例えば、語義別に分類された用例文集合が存在すれば、語義ごとに周辺の共起語を分析することで語義識別モデルを作成し、単語の意味を特定するための分類器を作ることができる。また、動詞についての格フレームを容易に自動構築することや語義ごとに項目分けをしたシソーラスを容易に構築することなどが可能となる。このようなシソーラスを構築するためには、単語に対する既存の語義識別能力を更に向上させることが不可欠である。単語が辞書中のどの意味区分に該当するのかを高い精度で識別することができれば、語義識別モデルを構築することに向けた学習データとしての利用や、意味を調べたい利用者に分かりやすい用例文を提供することへの利用などが可能となる。

語義識別システムは一般的に分類問題として定式化され、教師あり学習手法が用いられる。正解の語義が割り振られた用例文集合を教師データとし、その集合より語義を識別する分類モデルを構築する。この識別モデルに対して語義が不明な用例文を与え、各語義の中で最も相応しい語義を自動的に選択する。このとき、単語と共起する特徴を比較可能な形式に変換するために、頻度などを要素とするベクトルとして表現する。これにより、Support Vector Machine(SVM) [1] などといった教師あり学習手法を利用することが可能となる。

本稿では、既存の語義識別手法に対して更なる識別精度の改善を目的とするために、用例間距離学習手法を利用した語義識別モデルの構築について検討を行う。一般的にベクトル空間モデルを基本とした語義識別は、ある単語について同じ語義を持つ場合にはその単語の周辺において共起する単語の出現傾向が類似していると言われる。また、異なる語義で単語を使う場合には、一方の語義と比較して異なる単語が出現する傾向にある。距離学習手法は同じ語義を持つ特徴ベクトルの点集合は近い場所に集め、異なる語義を持つ点は遠い場所に離すことで、より語義識別しやすい特徴ベクトルを獲得する。

今回の報告では、最適な位置関係を得るために座標軸を変換する距離学習手法である Local Fisher Discriminant Analysis(LFDA)[3][4]、Semi-Supervised Local Fisher Discriminant Analysis(SELF) [5] を利用する場合と、データの移動を行いデータ間の最適な位置関係を求める距離学習手法である Neighborhood Component Analysis(NCA) と Large Margin Nearest Neighbor(LMNN) を利用する場合について語義識別実験を行った結果を示す。

2　ラベルによるデータ間距離の学習手法

教師データによる距離学習手法は、ラベル付きデータ集合に対して各データ間の距離をラベルに応じて変化させ、データ集合の最適な位置を求めるものである。同じラベルを持つデータ間の距離は短く、異なるラベルを持つデータ間の距離は遠くなるように変換を行う。

その際、距離学習の方法には大きく分けて、座標軸変換による学習とデータ移動による学習という2つの種類が存在する。本節ではこれら2つの学習手法の概要について説明する。

2.1　座標軸変換による距離学習

距離学習の方法で座標軸変換を利用することは、データ分析などでは一般的な方法としてよく利用される。これはラベル間の関係を調整するために、各ラベルに対してラベル内分散が最小、ラベル間分散が最大となるように、座標軸を回転させて最適なデータの位置関係を求める手法である。この考え方を利用した分

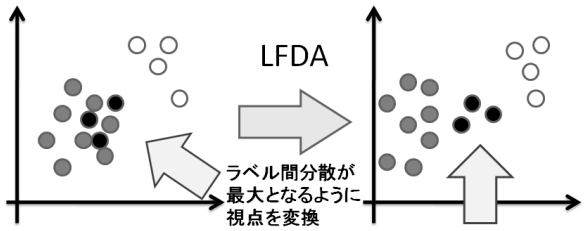

図 1: Local Fisher Discriminant Analysis

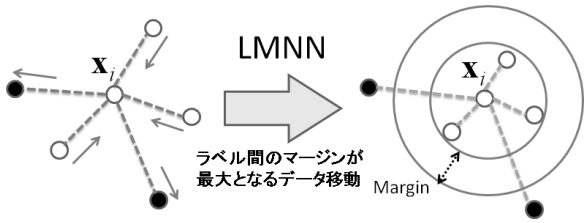

図 2: Large Margin Nearest Neighbor

析手法で代表的なものは、Local Fisher Discriminant Analysis(LFDA) (図 1) である [3][4]。LFDA ではスパースな行列に対して一般化固有値を計算することができない場合があるため、主成分分析を組み合わせた Semi-Supervised Local Fisher Discriminant Analysis(SELF) も存在する [5]。

この手法はデータの可視化や分析をする場合において、全データの位置関係を調べるときに有効な手段となる。しかし、この手法を利用して未知データの識別を行う場合は問題が生じる。ラベルに応じてデータが移動する訳ではなく、座標軸が回転されているため、SVM などで識別平面を求めると、同じ形の識別平面が回転して存在することになる。これにより、未知データを識別しても精度はほとんど変化しない結果となる[1]。従って、未知データに対してラベルの識別を行う際には、座標軸変換による距離学習と SVM などの識別平面による分類手法との組合せは適していない事がわかる。

2.2 データ移動による距離学習

距離学習の別の方法として、データ移動による手法も存在する。これは、座標軸を回転させてラベル間のデータ関係を最もよく表現する変換を行うのではなく、データそのものをラベルに応じて移動させることで最適なデータの位置関係を求める手法である。この考え方を利用した分析手法として、Neighborhood Component Analysis(NCA) [2] と Large Margin Nearest Neighbor(LMNN) [6] が存在する。

これらの手法は共にデータ間のマハラノビス距離を最適化するもので、それぞれの手法において設定した目的関数に対して最適な変換行列を求める。例えば、n 個の D 次元ベクトル $\mathbf{x}_i (i = 1, \cdots, n)$ と各ベクトルに対応するラベル $c_i (i = 1, \cdots, n)$ を考えたとき、2 つのベクトル $\mathbf{x}_i$ と $\mathbf{x}_j$ のマハラノビス距離は $d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{A}\mathbf{x}_i \quad \mathbf{A}\mathbf{x}_j)^T(\mathbf{A}\mathbf{x}_i \quad \mathbf{A}\mathbf{x}_j) = (\mathbf{x}_i \quad \mathbf{x}_j)^T\mathbf{M}(\mathbf{x}_i \quad \mathbf{x}_j)$ となる。ここで、行列 $\mathbf{M}$ は、$\mathbf{M} = \mathbf{A}^T\mathbf{A}$ を表し、これらの距離学習手法はこの行列 $\mathbf{M}$ を求めることが目的である。

[1]座標軸の回転をする際に次元縮退が同時に行われるため、その分に対応する少しの精度変化は存在する。

2.2.1 NCA の目的関数

NCA は 2 つのデータ $\mathbf{x}_i$ と $\mathbf{x}_j$ の近さを表す尺度 p_{ij} を

$$p_{ij} = \frac{\exp(\quad \|\mathbf{A}\mathbf{x}_i \quad \mathbf{A}\mathbf{x}_j\|^2)}{\sum_{k \neq i} \exp(\quad \|\mathbf{A}\mathbf{x}_i \quad \mathbf{A}\mathbf{x}_k\|^2)} \tag{1}$$

と表し、データ $\mathbf{x}_i$ に対して同じラベルを持つデータについて総和を求めたものが $\mathbf{x}_i$ の重要度となる。

$$p_i = \sum_{j \in C_i} p_{ij}, \quad (C_i = \{j | c_i = c_j\}) \tag{2}$$

NCA の目的関数は、この重要度 p_i をすべてのデータについての和を最大化することで、最終的にそのときの変換行列 $\mathbf{A}$ を求める。しかし、この目的関数は局所解に収束する可能性があるため、探索を行って収束したとしてもそれが大域的な最適解ではない場合がある。

2.2.2 LMNN

LMNN は図 2 に示すように、データ $\mathbf{x}_i$ に近い指定した数の同じラベルのデータは近くに移動し、異なるラベルのデータはマージンが最大となるように移動する。このとき、近傍に存在するデータを表すフラグ行列 η を定義し、データ $\mathbf{x}_j$ が $\mathbf{x}_i$ の近傍にある場合に $\eta_{ij} = 1$、近傍にない場合は $\eta_{ij} = 0$ とする。このとき、目的関数となるコスト関数は以下のように定義され、この関数を最小とする変換行列 $\mathbf{A}$ を求める。

$$\epsilon(A) = \sum_{ij} \eta_{ij} \|\mathbf{A}\mathbf{x}_i \quad \mathbf{A}\mathbf{x}_j\|^2 + c \sum_{ijl} \eta_{ij}(1 \quad \eta_{il}) [1 + \|\mathbf{A}\mathbf{x}_i \quad \mathbf{A}\mathbf{x}_j\|^2 \quad \|\mathbf{A}\mathbf{x}_i \quad \mathbf{A}\mathbf{x}_l\|^2]_+ \tag{3}$$

このコスト関数の第 1 項は同じラベルについての距離関係を表し、第 2 項は異なるラベルについての距離関係を表している。この関数を半正定値計画問題として最適解を求める。

3 距離学習を用いた語義識別手法

前節において紹介した距離学習手法を用いて語義識別を行う概要を説明する。

3.1 特徴抽出

語義の判別を行う単語を含む一文に対して、それと共起する単語を抽出する。本稿における語義識別手法では、学習データ、テストデータ共に単語として名詞と動詞を形態素解析を利用して抽出することとする。この共起単語についての頻度を要素とするベクトルを作成し、距離学習と語義識別に使用する。

3.2 距離学習とモデル構築

学習データに対して、距離学習手法を利用して語義識別モデルを構築する。本稿では、NCA、および、LMNN を利用して距離学習を行い、語義識別モデルに適用するためのデータに変換する。変換されたデータ集合に対して、NCA では SVM を利用して識別平面を求め、語義識別を行うためのモデルを構築する。また、LMNN では最近傍法を利用して、テストデータに最も近い学習データのラベルを判定結果として出力する。

3.3 語義の識別

構築した識別モデルに対して、語義を調べたいテストデータを入力し、自動的に語義の識別を行う。このとき語義の数が 3 個以上存在する場合は、SVM と LMNN では識別方法が異なる。SVM を利用する場合は、one-versus-rest 方式で各語義について繰り返し識別を行い、語義の識別をする必要がある。LMNN の場合は、One Nearest Neighbor(1-NN) 方式で、最も近い学習データの語義を識別結果とするため、繰り返し識別する必要はない。

4 実験

NCA、LMNN などの距離学習手法を利用した語義識別手法の精度を評価するために識別実験を行った。本節では、語義識別実験の概要を説明する。

4.1 データ

本実験で使用するデータは、Semeval2010 日本語 WSD タスクで課題として公開されたデータを利用する。これは 50 語の対象単語が指定され、その各単語についてそれを含む文を共起データとして使用する。共起データである文の数は学習データ、テストデータにおいて各 50 文用意され、学習データには対象単語の語義ラベルが付与されている。

4.2 評価方法

テストデータに対する語義識別結果を評価するために、50 件のデータに対する正解数を距離学習を行わず SVM で識別、NCA で距離学習し SVM で識別、LMNN で距離学習し 1-NN で識別した各正解数の比較を行う。また、各単語の正解数の比較だけではなく、全テストデータにおける各手法の正解率を平均的な精度として評価を行う。

5 実験結果と考察

5.1 テストデータによる識別

各手法に対する実験結果を表 1 に示す。NCA を利用した場合は、9 単語について精度が向上したものの、10 単語は精度が下がり、残りの 31 単語は変化なしの結果となった。全体的には性能改善の傾向が見られず、更なる改良が必要な結果となった。その方法として、学習データ用例文数の拡充、特徴抽出手法の改善、および、射影する次元数の最適化が考えられる。

LMNN を利用した場合は、SVM のみを利用する場合と比較して、精度が 68.9% から 69.6% と若干向上する結果が得られた。これより、NCA や LFDA、SELF を利用するよりも高い精度で識別可能なモデルの構築をすることができると考えられる。また、NCA では少ない学習データで距離学習を行っていたために局所解に収束し、識別精度が下がる傾向があったが、LMNN を利用し大域解を得るための変換行列を求めることで、識別精度が向上することも確認することが可能である。

表 1: 実験結果

単語	SVM	SELF+ SVM	NCA+ SVM	LMNN+ 1NN
現場	39	39	37	29
場所	48	48	48	48
取る	13	13	13	14
乗る	25	25	20	27
会う	33	33	33	33
前	31	31	29	27
子供	18	18	21	26
関係	39	39	39	39
教える	9	9	9	13
勧める	16	16	16	27
社会	43	43	43	42
する	21	21	23	20
電話	28	28	35	33
やる	47	47	47	47
意味	27	27	23	26
あげる	18	18	18	17
出す	14	14	17	26
生きる	47	47	47	47
経済	49	49	49	49
良い	12	12	15	23
他	50	50	50	50
開く	45	45	45	45
もの	44	44	44	44
強い	46	46	46	45
求める	38	38	38	39
技術	42	42	42	41
与える	29	29	28	25
市場	35	35	34	20
立つ	26	26	22	16
手	39	39	39	40
考える	49	49	49	49
見える	26	26	23	23
一	46	46	46	46
入れる	36	36	36	34
場合	43	43	43	45
早い	26	26	27	28
出る	30	30	30	28
入る	25	25	26	34
はじめ	30	30	33	44
情報	40	42	37	32
大きい	47	47	47	47
見る	40	40	40	40
可能	28	28	28	30
持つ	34	34	34	29
時間	44	44	42	44
文化	49	49	49	49
始める	39	39	40	39
認める	35	35	35	39
相手	41	41	41	40
高い	43	43	43	43
適合率	0.6888	0.6896	0.6876	0.6964

5.2 距離学習の効果

従来法としてよく使われる SVM に基づく語義識別ではデータ間の関連性などといったより深い分析作業に手間がかかる。しかし、距離学習に基づく語義識別ではこの作業を簡単に分析することが可能となる。まず、SVM とは異なり、1-NN を利用することでテストデータに対して最も近い学習データを特定することができる。テストデータに対して、最も近い学習データの選ばれる傾向を分析した結果、LMNN を利用した場合は 3 つ程度の特定の学習データのみで語義を識別する傾向があった。その中には単語数の少ない短い文が選ばれることが多かったが、どのような内容の文が識別に使われやすいのかなど、より深い分析は今後の課題として進めていく予定である。

また、SVM では識別する場合は、one-versus-rest 方式で繰り返し識別が行われる。このとき、3 つ以上語義がある場合は、テストテータと各ラベルの最短距離を比較することが難しい。LMNN では各ラベルとの最短距離を計算することが可能であるため、テストデータの識別しやすさを分析するには非常に有効な手段となる。また、新語義とみなされるデータの位置関係を調査する際の手段としても有効であると考えられる。

6 おわりに

本稿では、既存の語義識別手法に対して更なる識別精度の改善を目的とするために、用例間距離学習手法を利用した分類モデルの構築について検討した。その結果、LMNN を利用した語義識別手法を利用することで、従来よく利用される SVM よりも高い精度で識別することが可能であることを示した。また、LMNN を利用した場合は、3 つ程度の特定の学習データのみで語義を識別する傾向や 3 つ以上の語義を持つ場合の各語義間の関係を調べる上で有効な手段であることが分かった。今後の課題としては、教師データを利用した座標軸変換のより効果的な利用方法を考え、語義識別性能の改善を行う予定である。

参考文献

[1] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[2] Jacob Goldberger, Sam Roweis, Geoff Hinton, and Ruslan Salakhutdinov. Neighborhood Component Analysis. In *Proceedings of Advances of Neural Information Processing*, 2005.

[3] Masashi Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 905–912, New York, NY, USA, 2006. ACM.

[4] Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *J. Mach. Learn. Res.*, 8:1027–1061, May 2007.

[5] Masashi Sugiyama, Tsuyoshi Idé, Shinichi Nakajima, and Jun Sese. Semi-supervised local fisher discriminant analysis for dimensionality reduction. *Mach. Learn.*, 78:35–61, January 2010.

[6] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244, June 2009.