サポートベクターマシンに基づく Hit Miss Network を用いたインスタンス選択

小幡智裕*1　佐々木稔*2　新納浩幸*2
*1 茨城大学大学院理工学研究科情報工学専攻
*2 茨城大学工学部情報工学科

1　はじめに

機械学習の手法には大きく分けて、「教師あり学習」と「教師なし学習」がある。どちらも入力データが与えられたとき、それに対しての出力を正しく予測することが目的である。「教師あり学習」に限って述べると、入力に対して何を出力するべきであるか、学習データと呼ばれる入出力ペアの事例が複数与えられる。その学習データをもとに、新しいデータに対して適切なクラスをする手法のことである。

「教師あり学習」を用いた識別手法には、有名なもので k-NN(k-nearest neighbor algorithm) や SVM(support vector machine)といったものがあり、これらの手法は高い識別能力がある。しかしながら、学習データが増えると計算が膨大になるという問題点がある。

この問題点を解決するために、本研究ではサポートベクターマシンに基づく Hit Miss Network(HMN) を用いたインスタンス選択システムを提案し、実験及び評価を行う。このシステムにより得られる学習データをもとに、教師あり学習における学習データの質の向上を目指す。

評価実験ではいくつかの異なる評価用データセットを用意し、学習データのインスタンスを提案システムにより重要なものを残す。それを新しい学習データとして分類器を構築し、テストデータを識別する教師あり学習問題として評価を行う。従来手法では 1-NN を繰り返し利用してインスタンスの選択が行われているが[1]、提案システムでは SVM(support vector machine)を利用し、より優れた識別精度が得られるインスタンス選択手法を開発することが目的である。

提案システムを評価する際には、インスタンス選択により残った学習データの数を比較評価する他に、それを利用してテストデータが正しく識別された割合を求めた正解率も評価尺度とする。これを利用して、提案システムが従来手法と比較してどれだけ識別精度が向上するかを検証する。

2　学習手法

2.1 k-NN 法

k-NN 法は教師あり学習のひとつで、機械学習アルゴリズムの中で最も単純な手法と言われている。図 1 に k-NN 法の概念図を示す。

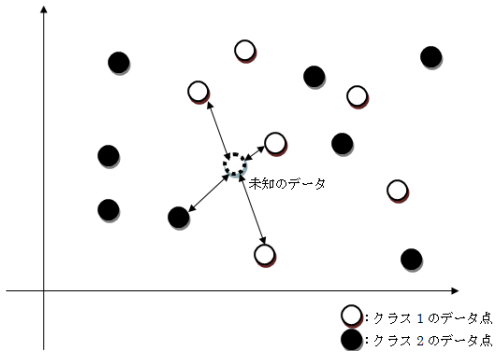

図 1：k-NN 法の概念図

分類したい未知データの周辺に存在する学習データとの距離を算出し、そのうち距離の最も近い k 個のデータ点の中から多数決でクラスを決めるものである。k=3 以上の場合では、周辺 3 つのデータ点のうち多数決で最も多いクラスに分類する。図 1 の例では、k=4 であり、多数決でテストデータはクラス 1 に分類される。また、k=1 の場合は、テストデータは最近接のデータ点のクラスに分類され、この手法は、1-NN(one nearest neighbor algorithm)と言われ、本研究ではこの 1-NN の手法を使用している。本研究では、ユークリッド距離を用いてデータ点間の距離を算出する。ユークリッド距離はデータ点 X, Y の座標をそれぞれ $x=(x_1,x_2,...,x_n)$, $y=(y_1,y_2,...,y_n)$ としたとき、2 点間の距離 $d(X,Y)$ を

$$d(X,Y)=\sqrt{\sum_{i=1}^{n}(y_i-x_i)^2}$$

と定義するものである。k-NN 法は、単純な操作ながら比較的高い精度を保っていて長く機械学習アルゴリズムとして使われて

いる。しかし一方で、外れ値に影響されやすいことや、データ量が大きくなると計算量が膨大になるという問題点もある。

2.2 SVM

SVM は線形識別器の教師あり学習アルゴリズムの 1 つであり、現在知られている多くの手法の中で最も識別能力が優れているものの 1 つとされている。識別器とは、ある学習データを 2 つのクラスのいずれかに識別することを目的としている。図 2 に SVM の概念図を示す。

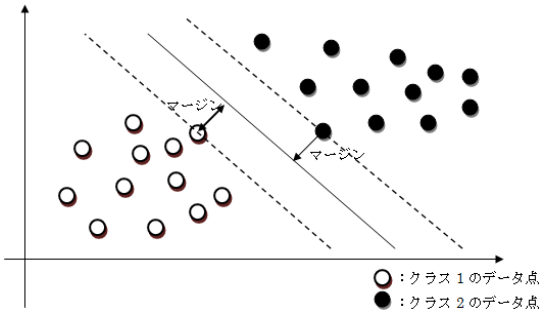

図 2：SVM の概念図

学習データに含まれる各データ点を $\{X_1, X_2, ..., X_n\}$ とする。また、それに対応するクラスを $c = (c_1, c_2, ...c_n)^T, c = \pm 1$ とする。このような学習データが与えられた場合、SVM は c=1 である点の集合と、c=-1 である集合とを上図のように分離する超平面を探し出すことを目標としている。SVM はまず、線形分離不可能な問題にも適用できるように、学習データを高次元特徴空間に写像する。そして、特徴空間上で学習データの中で最も他クラスと近い位置にあるデータ点を基準にし、そのユークリッド距離が最も大きくなるような位置に分離超平面を設定する。つまり、分離超平面と、2 種類のデータとの間の距離（マージン）が最大になるようにする。この分離超平面をもとに、未知のデータ点が属する側に対応するクラスを分類結果として出力される。

一般的にデータの特徴量を表わす次元を増やすと識別精度が悪くなるといわれているが、SVM はマージン最大化の概念によりデータの特徴量の次元が大きくなっても識別精度が悪くならないという利点がある。

逆に、学習データが増えると計算量が膨大になるということ、また多値分類問題への適応が難しいという欠点がある。

3 HMN を用いたインスタンス選択

3.1 HMN(Hit Miss Network)

HMN は、ベクトル情報とクラス情報を保持した学習データにおいて、各データ点と最近接にあるデータ点との関係を図式化して表現したものである。HMN を用いて各データ点の情報を読み取ることで、学習データ内のどのデータ点がクラス間の境界の点となっているかを算出することができる。HMN の例を図 3 に示す。

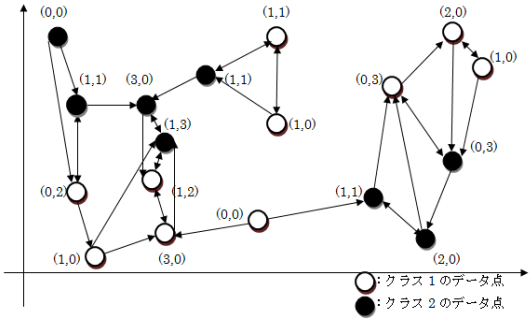

図 3：HMN の例

各データ点において、同一クラスで最近接である他の点に向けて、また異なるクラスで最近接である他の点に向けて、それぞれ hit, miss の関係を持つ有向辺を引く。ひとつのデータ点が他の点に向けた hit、miss の関係はそれぞれひとつしか存在しないが、他の点から受ける hit, miss の数はひとつのみとは限らない。

図 3 中の各データ点における数字は、(hit 次数,miss 次数)を表わしている。ここで、hit 次数とは他の点からの hit 辺の数、miss 次数とは他の点からの miss 辺の数のことである。また、hit 次数と miss 次数を合わせたものを総次数という。

一般的に、総次数が 0 のものは 1-NN においてクラス間の境界から遠い点であるとされる。そして、miss 次数が大きければ大きいほど 1-NN におけるクラス間の境界に近い点であるとされている[1]。

3.2 Hit Miss Score

Hit Miss score は次のように定義される関数の値である。

$$HMscore(p_a^h, p_a^m) = p_a^h \log \frac{p_a^h}{1/2\, p_a^h + 1/2\, p_a^m} - p_a^m \log \frac{p_a^m}{1/2\, p_a^h + 1/2\, p_a^m}$$

この式にある a は学習データ中のインスタンスで、$p^h{}_a$、$p^h{}_b$ はそれぞれ hit 次数、miss 次数を a における HMN の総次数で割

ったものである。HMscore の値が大きいインスタンスを学習データから取り除くと、SVM のような、マージン最大化アルゴリズムにおいて、他の多くの点のマージンが減少してしまうとされている[1]。

HMN における各データ点の次数と、HMscore には以下の関係がある。

- hit 次数の値が 0 のインスタンスは HMscore の値は 0 もしくは負となる。
- miss 次数の値が 0 かつ hit 次数が 0 より大きいインスタンスは HMscore の値が正となる。
- miss 次数よりも hit 次数が大きいインスタンスは HMscore の値が正となり、hit 次数より miss 次数が大きいインスタンスは HMscore の値が負となる。

3.3 HMN を用いたインスタンス選択アルゴリズム

はじめに、以下の手順であらかじめ学習データから重要ではないデータを取り除く。

1. 学習データの各データ点について、HMN を算出する
2. 各データ点の HMscore を算出する
3. HMscore の値が負のもの、または 0 であるデータ点は、学習データから取り除いても他のデータ点とのマージン減少の影響が小さいと判断し、学習データから取り除く
4. HMscore の値が正のデータ点の集合を S として出力する

3.4 1-NN に基づくインスタンス選択

従来手法は、3.3 節のアルゴリズムにより出力されたデータ集合の中から 1-NN に基づいたインスタンス選択を行い、最終的に学習データとなるインスタンスを決定する。従来手法のアルゴリズムを以下に示す。

1. 集合 S を入力とし、S の各データ点の HMN を構成する
2. Miss 次数が 0 より大きいデータ点の集合を S_0 とし、それ以外のデータ点の集合を S_1 とする
3. S_1 の HMN を算出し、S の HMN の総次数が 0 でないデータ点で、かつ S_1 の HMN の miss 次数が 0 より大きいデータ集合を S_t とする
4. $S_0 \cup S_t$ と S_0 について、leave-one-out による 1-NN を適用し、エラー数による比較を行う
5. S_0 の精度がよければ S_0 を新しい学習データとして出力し、$S_0 \cup S_t$ の精度がよければ $S_0 \cup S_t$ を S_0 として更新し、2. へ戻る

このアルゴリズムは、HMN の理論をもとに識別精度が落ちないようにインスタンスを選択し、学習データの数を可能な限り減らすものである。ここで、leave-one-out とは、学習データから 1 つのデータを抽出し、それをテストデータとして分類を行い、精度を評価する手法である。

3.5 SVM に基づくインスタンス選択アルゴリズム

従来手法では集合 $S_0 \cup S_t$ と集合 S_0 エラー数の比較に 1-NN を用いたが、提案手法ではこの部分における比較に SVM を用いることでインスタンス選択の精度向上を図る。以下に提案手法のアルゴリズムを示す。

1. 集合 S を入力とし、S の各データ点の HMN を算出する
2. Miss 次数が 0 より大きいデータ点の集合を S_0 とし、それ以外のデータ点の集合を S_1 とする
3. S_1 の HMN を算出し、S の HMN の総次数が 0 でないデータ点かつ、S_1 の HMN の miss 次数が 0 より大きいデータを S_t とする
4. $S_0 \cup S_t$ と S_0 について、leave-one-out による SVM を適用し、エラー数による比較を行う
5. S_0 の精度がよければ S_0 を新しい学習データとして出力し、$S_0 \cup S_t$ の精度がよければ $S_0 \cup S_t$ を S_0 として更新し、2. へ戻る

HMN により選択された学習データの候補が最終的な学習データに取り入れられるか否かは、$S_0 \cup S_t$ と S_0 のエラー数の比較により決まる。よって、インスタンス選択後の新しい学習データの識別精度は 1-NN の識別精度に依存すると考えられる。そこで、1-NN による識別によって算出されたエラーを用いず、高い識別精度を誇る SVM を用いた識別精度による比較を行って新しい学習データを算出したほうがより良い結果が得られると考える。

4 実験

4.1 実験方法

実験で使用したデータの概要を以下に

表 1 として示す。

表 1：使用したデータ

データ名	文書数	語彙数	クラス数
text1	1946	7511	2
re1	1657	3758	25

上記のデータのうち、text1 については、ランダムに 250 文書を選び、200 文書をインスタンス選択前の初期の学習データとし、50 文書をテストデータとして使用した。re1 については、25 クラスあるうちの 2 つのクラスのデータを選び、そこからランダムに 250 文書を選び、200 文書をインスタンス選択前の初期の学習データとし、50 文書をテストデータとする。このような学習データとテストデータの様々な組み合わせを 20 パターンずつ用意し、実験を行った。

この実験における評価指標には、テストデータの識別を行った結果、テストデータの総数に対して正しく識別された割合を正解率として用い、最大値、最小値、平均値を計算した。

実験には以下の 3 種類のシステムを利用して制度の比較を行う。

I. 初期の学習データを用いた識別
II. 従来手法により選択された学習データを用いた識別
III. 提案手法により選択された学習データを用いた識別

4.2 実験結果

以下の表 2 に、分類実験の正解率、表 3 にインスタンス選択によるデータ数の変化を示す。

表 2：実験結果

最大値			
データ名	I	II	III
text1	98	**96**	**96**
re1	68	64	**66**
最小値			
データ名	I	II	III
text1	84	68	**72**
re1	38	32	**44**
平均値			
データ名	I	II	III
text1	92.2	82	**84.1**
re1	53.7	49.9	**51.8**

5 考察

表 2 より、従来手法と提案手法の結果を比較すると、最大値、最小値、平均値のいずれにおいても、従来手法より提案手法の識別精度が高いという結果が得られたことが読み取れる。これは、5 章で述べたように提案手法のアルゴリズムにおいて、学習データに取り入れるインスタンスを選択する際の基準が重要であり、1-NN を基準としたときよりも SVM を基準としたときのほうがより良い識別精度を実現できたと考える。また、表 3 より、インスタンス選択により学習データのデータ数が大幅に削減されていることが読み取れる。最大値、最小値に関してはばらつきがあるものの、平均値で見ると従来手法より提案手法のデータ数のほうが少ない。

表 3：データ数

最大値			
データ名	I	II	III
text1	200	72	**62**
re1	200	**11**	13
最小値			
データ名	I	II	III
text1	200	**37**	38
re1	200	6	**5**
平均値			
データ名	I	II	III
text1	200	56.6	**51.7**
re1	200	8.25	**7.95**

これらのことから、提案手法は従来手法に比べ、分離平面に近い重要な点を残し、分離平面から遠い点を取り除くことができていると考える。

6 おわりに

本稿では、サポートベクターマシンに基づく Hit Miss Network(HMN) を用いたインスタンス選択システムを提案した。実験結果から、提案手法は従来手法に比べ、分類精度を下げることなく分離平面に近い重要な点を残すことが分かった。

今後の課題としては、提案手法はインスタンス選択を行う際にかかる時間が大きいためこれを改善することが課題である。

参考文献

[1] E.Marchiori. Class Conditional Nearest Neighbor for Large Mergin Instance Selection. *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol.32, no. 2, pp. 364-370, Feb.2010.