6C-2

自己増殖型ニューラルネットワークを用いたオンラインプロトタイプ抽出手法

神谷　祐樹†　　石井　利明†　　長谷川　修‡
東京工業大学　大学院総合理工学研究科　知能システム科学専攻†
東京工業大学　大学院理工学研究科　像情報工学研究施設‡

１．はじめに

最近傍識別は、（1）学習データが十分にある場合、ベイズ誤り確率の二倍未満の誤り確率を達成可能、（2）学習データを容易に追加可能、（3）kernel関数などの事前知識を必要としない、（4）最適化処理や識別関数のモデル化などを必要としない、などの利点がある。しかし、一般的に最近傍識別では全ての学習データをプロトタイプとして採用するため、大規模データを用いた際のメモリ使用量と識別速度が問題となる。

この問題の解決のために、少ないプロトタイプ数の識別器を構成することで、メモリの節約と識別速度の改善が図られる。代表的な手法として、k-Means Classifier (KMC) [1]、Learning Vector Quantization (LVQ) [2]、Nearest Subclass Classifier (NSC) [3]が挙げられるが、これらは追加学習やオンライン学習に対応していない。

そこで本稿では、オンライン学習可能なプロトタイプ抽出手法 k-SOINN について述べる。この手法によって、追加的な学習データを容易に学習し、プロトタイプ数の少ない最近傍識別器を構成可能となる。なお本稿では、全ての実験において特に断りの無い限り、識別手法に 1-NN 法を採用している。

２．提案手法

k-SOINNは、各クラスの入力に対して独立に自己増殖型ニューラルネットワーク（SOINN [4]）の第一層とk-means法[1]を用いる。本手法では、SOINNの第一層のみを採用し、手法を平易にすることで、SOINNの 5 つのユーザ定義パラメータを除いた。SOINN第一層は、ノイズに対して頑健に入力の分布のオンライン学習が可能である。しかし、SOINNは教師なし学習による分布の近似を目的としており、教師あり学習や高速な識別のためのプロトタイプ集合の抽出を考慮したものではない。そこで、識別境界の決定に重要な分布の端の領域を捉えるために、k-means法を用いた。

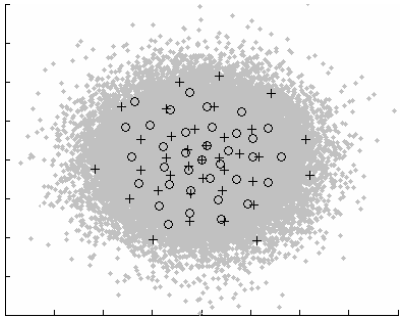

図１ k-SOINN と SOINN 第一層のみの学習結果の違い

k-means 法は、教師なしデータ集合からクラスタとそのクラスタの重心を探索する手法である。k-means 法では探索の際にクラスタ数とクラスタ重心の初期値を設定しなければならず、またそれらの値が結果に大きく影響するため、その値の調節が課題となる。本手法では、それらの初期値に SOINN 第一層の学習したノード数とその位置を設定することで、煩雑な調節過程を必要としない。また本手法では入力データをバッファリングし、一時的に保存した入力データ数が T となる毎に k-means 法を実行する。k-means 法を段階的に行うことで、オンライン学習の性質を損なわずに分布の端を捉えることができる。

k-SOINN と SOINN 第一層のみの学習結果の違いを図 1 に示す。この例では、正規分布からランダムに選択した入力サンプルを、k-SOINN および SOINN 第一層のみにオンライン学習させた。“＋”は k-SOINN の学習したノード位置、“○”は SOINN 第一層のみで学習したノード位置を表す。学習したノード数は同数であるが、k-SOINN のノード位置が分布の端に位置していることが確認できる。

３．評価実験と考察

提案手法の性能を評価するため、他手法との比較実験を行った。

Online Prototype-based Classifier Based on Self-organized Incremental Neural Network
†Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology
‡Imaging Science and Engineering Lab., Tokyo Institute of Technology

	Iris	Cancer	Ionosphere	Glass	Liver	Pima	Wine	Total
K-SOINN(X)	**96.67** ±1.15	96.49 ±0.0	86.78 ±0.99	65.43 ±1.84	63.19 ±1.46	68.96 ±1.71	72.33 ±0.35	78.55 ±1.07
	6.65 (3)	**0.33** (1)	**4.54** (8)	27.43 (11)	18.56 (36)	8.41 (22)	**1.87** (1)	**9.68**
KMC(M)	96.2 ±0.8	95.9 ±0.3	87.4 ±0.6	68.8 ±1.1	59.3 ±2.3	68.7 ±0.9	71.9 ±1.9	78.3 ±1.1
	8.0 (4)	**0.29** (1)	**4.0** (7)	**17** (6)	11 (19)	**1.0** (4)	29 (17)	**10.04**
NSC($\sigma^2_{\max}$)	96.3 ±0.4	**97.2** ±0.2	**91.9** ±0.8	70.2 ±1.5	62.9 ±2.3	68.6 ±1.6	**75.3** ±1.7	80.4 ±1.2
	7.3 (0.25)	1.8 (35.0)	31 (1.25)	97 (0.005)	**4.9** (600)	1.7 (2600)	96 (4.0)	34.24
LVQ(M)	96.1 ±0.6	96.3 ±0.4	86.4 ±0.8	68.3 ±2.0	66.3 ±1.9	73.5 ±0.9	72.3 ±1.5	79.9 ±1.2
	15 (22)	5.9 (40)	6.8 (24)	45 (97)	8.4 (29)	3.4 (26)	32 (57)	16.64
NNC (k)	**96.7** ±0.6	**97.0** ±0.2	86.1 ±0.7	**72.3** ±1.2	**67.3** ±1.6	**74.7** ±1.4	73.9 ±1.9	**81.14** ±1.09
	100 (14)	100 (5)	100 (2)	100 (1)	100 (14)	100 (17)	100 (1)	100

表 1 識別性能とプロトタイプ数の比較：上段は識別率(%)、下段はプロトタイプ使用率 r_c（括弧内は最適化されたパラメータの値）を表す。太字は各データセットに対する最高値およびそれに準じた値を表す。

比較手法には、NSC($\sigma^2_{\max}$) [3]、KMC(M) [1]、LVQ(M) [2]を用いた。本実験では、これらの手法の識別率とメモリ使用量および識別速度を k-最近傍識別器(NNC(k))と比較してその性能を評価した。本稿では、学習後のメモリの節約と識別速度の度合として、プロトタイプ使用率（r_c = プロトタイプ数/学習サンプル数）を定義し、識別率と r_c を識別器の性能の評価基準とした。

本実験では、UCI Repository [5]の Iris、Breast Cancer など 7 種類のデータセットを用い、各データセットに対する識別器の性能と平均値を検証した。各データセットのサンプル数および次元数は関連する文献[5]を参照されたい。比較手法の結果は Veenman らが行った同様の実験の結果（[3]、表 2）を用いた。比較手法のパラメータは cross-validation によって最適化されている。また、提案手法の 2 つのパラメータ（a_d、λ）はプロトタイプ数に影響するが、値の大小による影響のみで個々の値の違いによる影響は少ない。したがって本実験では 2 つのパラメータを同一の値 X とした（$k-SOINN(X)$、$X = a_d = \lambda$）。X の値は 10-fold cross-validation を行って決定した。

表 1 に k-SOINN と他手法との比較実験の結果を示す。表の各要素の上段は各データセットに対する識別率を表し、下段はプロトタイプ使用率 r_c を表す。識別率は 10-fold cross-validation による 10 回の試行の平均値と分散を示している。この結果から、NSC および LVQ は高い識別性能を示すが、プロトタイプ使用率の値が高いことから、識別に時間が掛かることが予想される。

また KMC は、識別性能は多少低いが、高速な識別が可能である。提案手法である k-SOINN は KMC と同程度の識別性能を示しており、全データセットの平均値において、比較した全手法中で最良の r_c 値を示している。KMC は各クラスのプロトタイプ数を同一定数に設定する必要があるのに対し、k-SOINN は各クラスに必要なプロトタイプ数を自動的に決定することができる。そのため、提案手法が KMC より少ないプロトタイプ数で同程度の性能を示したと考えられる。

また、提案手法のパラメータ値の決定基準を識別率重視にした場合には、全データセットの平均識別率は 79.83%（r_c=28.00）となり、NSC および LVQ と同等程度となった。この結果から、提案手法はタスクに応じて 2 つのパラメータを変更することで、識別器の性質を変化させることができる。

謝辞 本研究の一部は NEDO 産業技術研究事業助成金を受けて実施している。

参考文献

[1] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2001.

[2] T. Kohonen, “Improved Versions of Learning Vector Quantization," Proc. Int'l Joint Conf. Neural Networks, Vol. 1, pp. 545-550, 1990.

[3] C.J. Veenman and M.J.T. Reinders, “The Nearest Subclass Classifier: A Compromise between the Nearest Mean and Nearest Neighbor Classifier,” IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 27, No. 9, pp. 1417-1429, 2005.

[4] F. Shen and O. Hasegawa, “An Incremental Network for On-line Unsupervised Classification and Topology Learning,” Neural Networks, Vol. 19, pp. 90-106, 2006.

[5] C. Merz and M. Murphy, “UCI Repository of Machine Learning Databases,” Irvine, CA, University of California Department of Information, 1996. http://www.ics.uci.edu/~mlearn/MLRepository.html