Neural Networks 21 (2008) 1537–1547

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A fast nearest neighbor classifier based on self-organizing incremental neural network

Shen Furao [a,*], Osamu Hasegawa [b]

[a] The State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210093, PR China

[b] Imaging Science and Engineering Lab., Tokyo Institute of Technology, Japan

ARTICLE INFO

Article history:
Received 29 December 2007
Received in revised form
21 May 2008
Accepted 2 July 2008

Keywords:
Self-organizing incremental neural network
Nearest neighbor
Fast
Prototype-based classifier

ABSTRACT

A fast prototype-based nearest neighbor classifier is introduced. The proposed Adjusted SOINN Classifier (ASC) is based on SOINN (self-organizing incremental neural network), it automatically learns the number of prototypes needed to determine the decision boundary, and learns new information without destroying old learned information. It is robust to noisy training data, and it realizes very fast classification. In the experiment, we use some artificial datasets and real-world datasets to illustrate ASC. We also compare ASC with other prototype-based classifiers with regard to its classification error, compression ratio, and speed up ratio. The results show that ASC has the best performance and it is a very efficient classifier.

1. Introduction

k-nearest neighbors algorithm (Cover & Hart, 1967) is very useful for some applications such as machine learning, data mining, natural language understanding, and information retrieval (Dasarathy, 1991). Let $T = \{t_i \in \omega, i = 1, 2, \ldots, m\}$ denotes a set of training patterns, each pattern $t_i \in T$ has a class label. The target of kNN is to find the k-nearest neighbors of a test pattern x ($x \in \omega$) in T based on a dissimilarity measure $d(.,.)$, and then classify the pattern x with the same label as the majority voting of nearest patterns in the training set. We call this classifier as Nearest Neighbor Classifier (NNC(k)) (Cover & Hart, 1967). If we set $k = 1$, the k-nearest neighbors classifier becomes 1-nearest neighbor (1-NN) classifier.

The main advantages of kNN include that it can learn from a small set of examples, can incrementally add new information at run time, no optimization required, capable to model very complex target functions by a collection of less complex approximations, etc.; on the other hand, its major disadvantage is that it is computationally intensive for large datasets.

kNN uses all training data as the prototypes. To reduce the amount of required storage and improve the classification speed, we need to reduce the number of prototypes. The question becomes, for the training dataset $T = \{t_i \in \omega, i = 1, 2, \ldots, m\}$, to find a set P with M prototypes that represent T such that P can be used for classification using the nearest neighbor rule.

For the prototype-based algorithms, the question is how one knows when there are enough prototypes and how to prevent overfitting to training data. NNC uses all the training data to label unseen patterns. The Nearest Mean Classifier (NMC) (Hastie, Tibshirani, & Friedman, 2001) only stores the mean of each class, i.e. one prototype per class. It generally has a high error on the training and test data. There are lots of other prototype-based algorithms such as k-means classifier (KMC) (Hastie et al., 2001), Learning Vector Quantization (LVQ) (Kohonen, 1990), and others (Bezdek & Kuncheva, 2001; Wilson & Martinez, 2000). Such methods select a fixed number of prototypes per class as an overfitting avoidance strategy. But when the class distributions differ from each other, either in the number of patterns, the density of the patterns, or the shape of the classes, the optimal number of prototypes may be different for each other. Nearest Subclass Classifier (NSC) (Veenman & Reinders, 2005) tried to impose the number of prototypes per class by introducing a variance constraint parameter. They assume that the features of every pattern contain the same amount of noise and do not model label noise, and they assume the undersampling of the classes is the same everywhere in feature space. Such assumptions may not be satisfied in real world task.

For some prototype-based classifiers, it is very difficult to incrementally add new information at run time, and to eliminate the influence of noise. Here, noise means the unknown amount of noise in the features and class labels of the training dataset.

In this paper, we propose a prototype-based nearest neighbor method that is based on the self-organizing incremental neural

* Corresponding author. Tel.: +81 45 924 5180; fax: +81 45 924 5175.
E-mail address: frshen@nju.edu.cn (F. Shen).

0893-6080/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2008.07.001

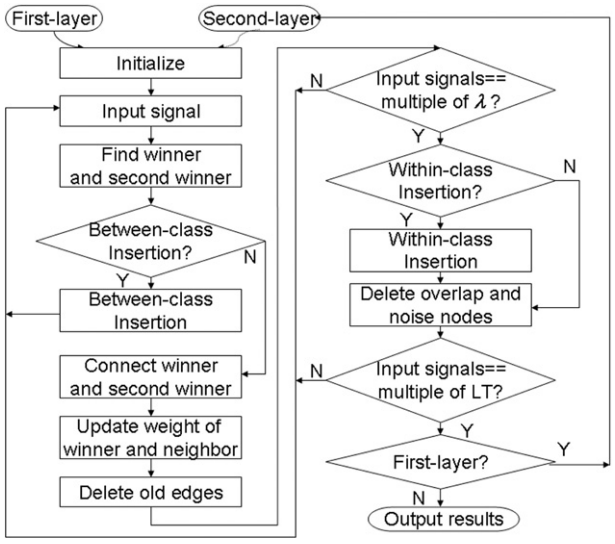

Fig. 1. Learning process of SOINN.

network (SOINN) (Shen & Hasegawa, 2006, 2005). The goals of the proposed method are: (1) automatically learn the number of prototypes needed to represent every class, if needed, allocate different number of prototypes for different classes with different distribution or shape; (2) learn new information without destroying old learned information, i.e., to realize incremental learning; (3) reduce prototypes caused by noise in order to decrease the misclassification, i.e., the proposed method must be robust to noisy training data; (4) delete unnecessary prototypes during the classification process to accelerate the classification speed, i.e. only the prototypes used to determine the decision boundary will be remained.

During the classification process, if not stated differently, for all experiments in this paper we will employ the 1-nearest neighbor (prototype) rule to classify patterns based on the generated set of labeled prototypes.

2. Overview of self-organizing incremental neural network (SOINN)

In this section, we introduce the self-organizing incremental neural network (SOINN) (Shen & Hasegawa, 2006), which is the basis of the proposed method. SOINN adopts two-layer network. The training results of first-layer will be used as the training set for second-layer. The targets of SOINN are realizing the unsupervised learning and represent the topology structure of input distribution. We summarize the flowchart of SOINN in Fig. 1.

When an input vector is given to SOINN, it finds the nearest node (winner) and the second nearest node (second winner) of the input vector. It subsequently judges if the input vector belongs to the same cluster of the winner or second winner using the similarity threshold criterion. The similarity threshold T_i is defined as the distance (Euclidean distance) from the boundary to the center of Voronoi region V_i of node i. During the learning process, the node i will change its position to meet the inputting pattern distribution, and thus the Voronoi region V_i of the node i will also change, therefore, the similarity threshold T_i will also change.

Algorithm 2.1. Calculation of similarity threshold T

(1) Initialize the similarity threshold of node i to $+\infty$ when node i is generated as a new node.
(2) When node i is a *winner* or *second winner*, update similarity threshold T_i:

- If the node has direct topological neighbors, T_i is updated as the maximum distance between node i and all of its neighbors,
$$T_i = \max_{c \in N_i} \|\boldsymbol{W}_i - \boldsymbol{W}_c\|, \quad (1)$$
here, N_i is the neighbor set of node i.
- If node i has no neighbor, T_i is updated as the minimum distance of node i and all other nodes in A,
$$T_i = \min_{c \in A \setminus \{i\}} \|\boldsymbol{W}_i - \boldsymbol{W}_c\| \quad (2)$$
here, A is the node set.

Here, *winner* means the nearest node to the input pattern, and *second winner* means the second nearest node to the input pattern. $\boldsymbol{W}_i$ is the weight vector of node i.

The input vector will be inserted to the network as a new node to represent the first node of a new class if the distance between the input vector and the winner or second winner is greater than the similarity threshold of a winner or second winner. This insertion is called a between-class insertion because this insertion will result in the generating of a new class, even if the generated new class might be classified to some older class in the future.

If the input vector is judged as belonging to the same cluster of winner or second winner, and if no edge connects the winner and second winner, connect the winner and second winner with an edge, and set the 'age' of the edge as '0'; subsequently, increase the age of all edges linked to the winner by '1'.

Then, update the weight vector of the winner and its neighboring nodes. We use i to mark the winner node, and M_i to show the times for node i to be a winner. The change to the weight of winner $\Delta \boldsymbol{W}_i$ and change to the weight of the neighbor node $j(\in N_i)$ of $i\Delta \boldsymbol{W}_j$ are defined as

$$\Delta \boldsymbol{W}_i = \frac{1}{M_i}(\boldsymbol{W}_s - \boldsymbol{W}_i) \quad (3)$$

and

$$\Delta \boldsymbol{W}_j = \frac{1}{100 M_i}(\boldsymbol{W}_s - \boldsymbol{W}_j) \quad (4)$$

where $\boldsymbol{W}_s$ is the weight of the input vector.

If the age of one edge is greater than a predefined parameter a_d, then remove that edge.

After λ learning iterations, the SOINN inserts new nodes into the position where the accumulating error is extremely large. Cancel the insertion if the insertion cannot decrease the error. The insertion here is called within-class insertion because the new inserted node is within the existing class; also, no new class will be generated during the insertion. Then SOINN finds the nodes whose neighbor is less than or equal to 1 and deletes such nodes based on the presumption that such nodes lie in the low-density area.

After LT learning iterations of the first layer, the learning results are used as the input for the second layer. The second layer of SOINN uses the same learning algorithm as the first layer.

SOINN adopts two schemes to insert new nodes and thus realize the incremental learning and topology representation: between-class insertion and within-class insertion. In order to realize the within-class insertion, SOINN uses 5 user determine parameters and another parameter "error-radius" to judge if the insertion is successful, which makes the system complicated and difficult to understand.

In fact, during the training of first-layer of SOINN, between-class insertion is the main part, and within-class insertion has little contribution for inserting new nodes. During the training of second-layer of SOINN, both between-class insertion and within-class insertion are needed to make the number of nodes enough for representing topology structure (Shen & Hasegawa, 2007).

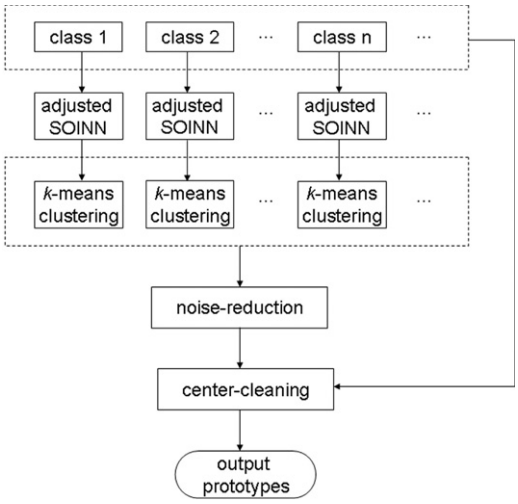

Fig. 2. Learning process of adjusted SOINN classifier (ASC).

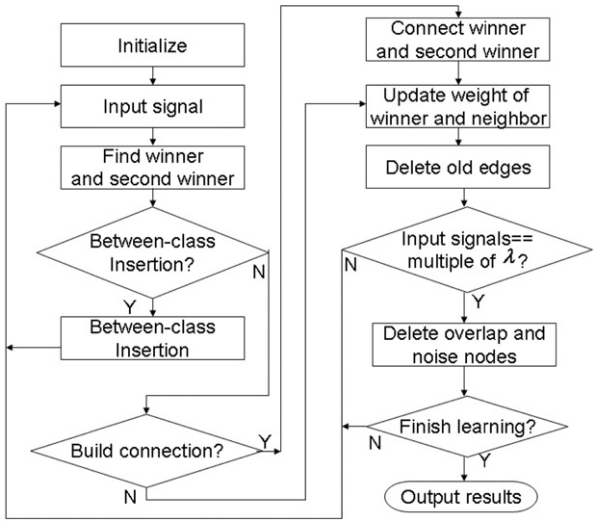

Fig. 3. Flowchart of adjusted SOINN.

3. Adjusted SOINN classifier (ASC)

The proposed Adjusted SOINN Classifier (ASC) inherited some properties of SOINN such as incremental learning, robust to noise, and automatically learn number of prototypes needed to represent every class. The shortcoming of SOINN is that it needs too much parameters to realize the within-class insertion; it is not stable and the results depend on the input sequence of the training data; also, the target of SOINN is to realize unsupervised learning and topology representation, and here we want to do supervised learning and use as less prototypes as possible to realize fast classification. We improve SOINN with the following aspects: (1) adjust SOINN with less parameters to represent the topology structure of input data; (2) improve SOINN to get more stable results with the help of *k*-means clustering (Hastie et al., 2001); (3) design the noise-reduction part to reduce some prototypes caused by noise; and (4) design the center-cleaning part to delete those unnecessary prototypes during the classification process. Because the proposed method is based on SOINN, we name the proposed algorithm Adjusted SOINN Classifier (ASC). The flowchart of ASC is shown in Fig. 2.

From Fig. 2 we know that, at first, ASC does adjusted SOINN for every class separately, then does *k*-means clustering with the results of adjusted SOINN for every class, then use all the *k*-means results to do noise-reduction. At last, ASC uses the input data of all classes and the results of noise-reduction part to do center-cleaning process.

3.1. Adjusted SOINN

As we pointed out in Section 2, during the training of first-layer of SOINN, between-class insertion is the main part, and within-class insertion has little contribution for inserting new nodes. The target of second-layer is to delete redundant nodes, separate overlapped clusters, and delete nodes caused by noise. Just for the topology representation target, the first-layer can get better topology representation than second-layer (Shen & Hasegawa, 2007). Here we only adopt the first-layer of SOINN as the basis of the proposed ASC method, and delete the within-class insertion part to make it easy to understand and save 5 user-determine-parameters. The deletion of within-class insertion will not influence the learning results. It is because if we only adopt single-layer network, the between-class insertion assures that the density of nodes will be enough to represent the topology structure. With the definition of similarity threshold, inserted new nodes may come from not happened area, and the following process (such as competitive Hebbian rule) will link such new nodes with old nodes. The adaptively updated similarity threshold will not be too large to make the nodes sparse. With between-class insertion, when the nodes reach the boundary of a class, the insertion will be automatically stopped for the class, and we avoid the permanent increase of nodes. Fig. 3 gives the flowchart of adjusted SOINN.

When an input vector is given to adjusted SOINN, it finds the winner and second winner of the input vector, then judges whether the input vector belongs to the same cluster of the winner or second winner using the criterion of similarity threshold. The similarity threshold T_i is calculated using Algorithm 2.1.

If the distance between the input vector and the winner or second winner is greater than the similarity threshold of winner or second winner, the input vector will be inserted to the network with between-class insertion process.

If the input vector is judged as belonging to the same cluster of winner or second winner, we update the weight vector of winner and its neighbor nodes with formula (3) and (4).

If no edge connects the winner and second winner, we connect the winner and second winner with an edge, and set the 'age' of the edge as '0'; subsequently, we increase the age of all edges linked to the winner by '1'. If the age of one edge is greater than a predefined parameter a_d, then we remove that edge.

After λ learning iterations, adjusted SOINN finds the nodes whose neighbor is less than or equal to 1 and deletes such nodes. Algorithm 3.1 is the detail algorithm of adjusted SOINN.

Algorithm 3.1. Adjusted SOINN

(1) Initialize node set A to contain two nodes, c_1 and c_2 with weight vectors chosen randomly from the input pattern. Initialize connection set C, $C \subset A \times A$, to the empty set.
(2) Input new pattern $\xi \in R^n$.
(3) Search the nearest node (*winner*) s_1, and the second-nearest node (*second winner*) s_2 by

$$s_1 = \arg\min_{c \in A} \|\xi - \boldsymbol{W}_c\| \tag{5}$$

$$s_2 = \arg\min_{c \in A \setminus \{s_1\}} \|\xi - \boldsymbol{W}_c\|. \tag{6}$$

If the distance between ξ and s_1 or s_2 is greater than similarity threshold T_{s_1} or T_{s_2}, the input signal is a new node, add it to A and go to step (2) to process the next signal. The similarity threshold T is calculated by Algorithm 2.1.

(4) If a connection between s_1 and s_2 does not exist, create it. Set the age of the connection between s_1 and s_2 to zero.
(5) Increment the age of all edges emanating from s_1 by 1.
(6) Adapt the weight vectors of the *winner* and its direct topological neighbors by fraction $\epsilon_1(t)$ and $\epsilon_2(t)$ of the total distance to the input signal,

$$\Delta \boldsymbol{W}_{s_1} = \epsilon_1(t)(\xi - \boldsymbol{W}_{s_1}) \tag{7}$$

and for all direct neighbors i of s_1,

$$\Delta \boldsymbol{W}_i = \epsilon_2(t)(\xi - \boldsymbol{W}_i). \tag{8}$$

We adopt a scheme to adapt the learning rate over time by

$$\epsilon_1(t) = \frac{1}{t} \tag{9}$$

$$\epsilon_2(t) = \frac{1}{100t}. \tag{10}$$

(7) Remove edges with an age greater than a predefined threshold a_d. If this results in nodes having no more emanating edges, remove them as well.
(8) If the number of input signals generated so far is an integer multiple of parameter λ, delete some nodes as follows: for all nodes in A, search for nodes having no neighbor or only one neighbor, then remove them.
(9) Go to Step (2) to continue the learning until the learning time is satisfied.

In Algorithm 3.1, we need to determine two parameters a_d and λ. These two parameters will influence the frequency of deleting connections between nodes and nodes lie in sparse area. Thus, if we want to save previous learned knowledge much longer, we choose large value for these two parameters, and get lots of nodes to realize low classification error; on the other hand, if we want less nodes to save memory space and speed up the classification, we set the value of these two parameters small to remove nodes and edges frequently. It means that, the two parameters are depending on the real condition of the task, we can use these parameters to control the recognition performance of ASC.

3.2. k-means clustering

k-means clustering (Hastie et al., 2001) is a method for finding clusters and cluster centers in a set of unlabeled data. We choose the desired number of cluster centers m, give an initial set of centers $c_j(0), j = 1, \ldots, m$, the k-means algorithm alternates the two steps: (1) for each center we identify the subset of training points (its cluster) that is closer to it than any other center; (2) the means of each feature for the data points in each cluster are computed, and this mean vector becomes the new center for that cluster. These two steps are iterated until convergence.

Algorithm 3.2. k-means clustering

(1) Partition the inputting vector data $x_i, i = 1, \ldots, n$ into the channel symbols using the minimum distance rule. This partitioning is stored in a $n \times m$ indicator matrix S whose elements are defined by

$$s_{ij} = \begin{cases} 1 & \text{if } d(x_i, c_j(k)) = \min\limits_p d(x_i, c_p(k)) \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

(2) Determine the centroids of the inputting data by channel symbol. Replace the old centers with these centroids:

$$c_j(k+1) = \frac{\sum\limits_{i=1}^{n} s_{ij} x_i}{\sum\limits_{i=1}^{n} s_{ij}}, \quad j = 1, \ldots, m. \tag{12}$$

(3) Repeat step (1)–step (2) until no $c_j, j = 1, \ldots, m$ changes anymore.

k-means heavily depends on the initial value of centers $c_j(0), j = 1, \ldots, m$. The difficulties of k-means are how to determine the number of centers m in advance and how to find good initial value of centers. To solve such difficulties, in ASC, we use the learned number of nodes of adjusted SOINN as the number of centers, and use the position (weight vector) of adjusted SOINN nodes as the initial value of centers.

In ASC, adjusted SOINN is not stable and the results depend on the sequence of input data. The number of nodes and the position of nodes are different if we repeat the training under same environment with different input sequence. With the help of k-means clustering, we move such nodes to the center of the clusters and improve the stability of ASC. It is because that the generated nodes of adjusted SOINN represent the topology of input data, such nodes are distributed near the centers of sub-clusters, and k-means clustering moves such nodes to the centers of sub-clusters.

3.3. Noise-reduction part

If there is noise in the training data, during the training of adjusted SOINN, some nodes will be generated by noise. We only adopt single-layer in adjusted SOINN. Therefore, we cannot remove all nodes generated by noise if there is plenty of noise. Here, noise means that the training dataset contains an unknown amount of noise in the features and class labels. To prevent the nearest neighbor rule from fitting to the noisy training data without restriction, we use the idea of an early method: k-Edited Neighbors Classifier (ENC) (Wilson, 1972) in ASC, i.e. if the label of a node differs from the label of majority voting of its k-neighbors, it is considered an outlier and the node is removed from the set of prototypes. Algorithm 3.3 is the detailed algorithm of the Noise-reduction part.

Algorithm 3.3. Noise-reduction

(1) For a prototype c in prototype set C, find k nearest prototypes of prototype c.
(2) Delete prototype c from the prototype set C if the major voting of k nearest prototypes has a different class label with prototype c.
(3) Repeat this process until all prototypes are processed.

In this noise-reduction part, we can use some methods such as cross-validation to tune the parameter k.

3.4. Center-cleaning part

The prototype set obtained using adjusted SOINN and k-means clustering can be used to represent the topology structure of the input data. However, during the classification process, some prototypes in the central part of a class might not be useful because, during the classification process, we use the one-nearest neighbor (prototype) rule to classify patterns based on the generated set of labeled prototypes, and only prototypes lie in the boundary can be used. We must delete those prototypes which lie in the central parts of classes to save memory space of storage for prototypes and to accelerate the classification speed. We designed Algorithm 3.4 to realize this target. Algorithm 3.4 is based on this idea: if a prototype of a class has never been the nearest prototype to other classes, the prototype lies in the central part of the class. Therefore, we delete it.

Algorithm 3.4. Center-cleaning process

(1) Suppose that there are n classes. Given class i, $i = 1, \ldots, n$, do the following steps.
(2) For all samples of other classes that differ from class i, find the nearest prototype of class i generated by adjusted SOINN and k-means clustering.

(3) Delete this prototype if a prototype of class i has never been the nearest prototype of other classes. Repeat this step until no prototype can be deleted.

After executing of Algorithm 3.4, the remaining prototypes are all useful prototypes for the one-nearest-neighbor rule. Removed prototypes according to the center-cleaning process have no usage for the classification process. We must mention that the cleaning rule is based on the training set; it is possible for this process to delete some useful prototypes for the testing set and leads to a loss of the recognition performance, but if the testing set obeys the same distribution as the training set, the removal will not decrease the efficiency and can accelerate the speed greatly.

3.5. ASC algorithm

With the analysis of Sections 3.2–3.4, we give the whole learning algorithm of adjusted SOINN classifier (ASC) in Algorithm 3.5.

Algorithm 3.5. Learning algorithm of ASC

(1) Suppose there are n classes, for every class, do adjusted SOINN (Algorithm 3.1), Algorithm 3.1 outputs the number of prototypes $\{N_i, i = 1, \ldots, n\}$ of every class, and give the weight vector of such prototypes $\{AS(c_j, i), j = 1, \ldots, N_i; i = 1, \ldots, n\}$.
(2) For every class, do k-means clustering (Algorithm 3.2). The number of centers for every class may be different, we adopts the N_i generated by step (1) as the number of centers for class i. The $\{AS(c_j, i), j = 1, \ldots, N_i\}$ will be the initial value of centers for class i. The results of this step will be $\{km(c_j, i), j = 1, \ldots, N_i; i = 1, \ldots, n\}$.
(3) For all prototypes in $\{km(c_j, i), j = 1, \ldots, N_i; i = 1, \ldots, n\}$, use Algorithm 3.3 to remove those prototypes caused by noise (noise reduction), and we get $\{NR(c_j, i), j = 1, \ldots, \hat{N}_i; i = 1, \ldots, n\}$, here $\hat{N}_i$ is the new number of prototypes of class i, it will be equal to or less than N_i.
(4) For $\{NR(c_j, i), j = 1, \ldots, \hat{N}_i; i = 1, \ldots, n\}$, with all samples of all classes, use Algorithm 3.4 to delete central prototypes of every class (center cleaning), which are not useful for classification processes, and we get final prototype set $\{Prototype(c_j, i), j = 1, \ldots, M_i; i = 1, \ldots, n\}$, M_i is the final number of prototypes of every class, it is equal to or less than $\hat{N}_i$, and $\sum_{i=1}^{n} M_i$ is the final total number of prototypes. Such prototypes will be used for classification of new objects.

In Algorithm 3.5, first, we execute the adjusted SOINN separately for every class; then we use the training results of adjusted SOINN (number of prototypes and the position of prototypes) to do k-means clustering; then we use the k-means results of all classes to reduce some prototypes caused by noise. Finally, we use the remaining prototypes and all training data to find unnecessary prototypes that lie in the central part of node distribution, then remove them to speed up the process.

In the above algorithm, there are two parameters (a_d, λ) needed by adjusted SOINN, and we also need one parameter k in step (3) to realize the noise-reduction process. We have given the discussion of such parameters in Sections 3.2 and 3.3.

For the prototype-based classifier, the classification error, storage requirements, and speed can be used to evaluate the performance of a classifier. In this paper, we measure the memory requirements with the compression ratio r_c of the trained classifier, and measure the comparison of classification speed with the speed up ratio $r_s(A, B)$ between trained classifiers A and B, where

$$r_c = \frac{\text{number of prototypes}}{\text{train data size}} \tag{13}$$

$$r_s(A, B) = \frac{\text{classification speed of classifier } A}{\text{classification speed of classifier } B} \tag{14}$$

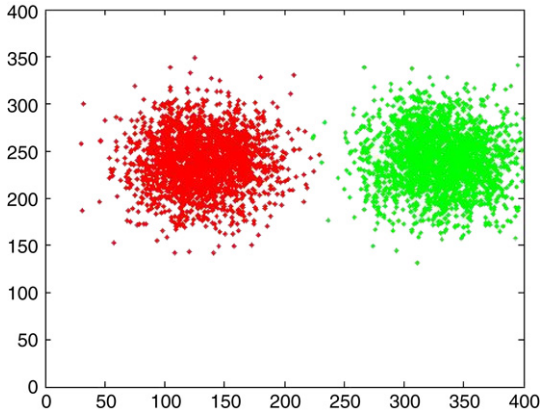

Fig. 4. Two Gaussian distribution datasets, without overlap (2 classes).

and use classification error, compression ratio, and speed up ratio to validate the performance of classifiers.

4. Experiment

In this part, at first, we use some 2-dimension artificial datasets to test the ASC and illustrator the detail of ASC; then through the test on a real-word dataset we prove the efficiency of the proposed method. At last we compare ASC with some other typical prototype-based classifiers.

In the following experiment, for software environment, the operating system is WinXP and the programming language is Visual C++6.0. Hardware uses a PC with an Intel Xeon(TM) CPU 3.20 GHz and 2.0 GB RAM.

4.1. Artificial dataset

During the test of all artificial datasets, we set the parameters as following: $a_d = 20$, $\lambda = 20$, $k = 5$. The a_d and λ will influence the number of generated prototypes, i.e. compression ratio, but they will not influence the recognition performance if we can get enough prototypes. The parameter k is not sensitive, we can choose other value for it and get the same recognition results. For all experiments in this section, we do 10 times learning and testing, and then give the average classification error and compression ratio as the recognition performance.

(1) Experiment 1: two Gaussian distribution without overlap

In this experiment, we adopt two non-overlapped Gaussian distribution (as shown in Fig. 4). For every class, we choose 2000 samples as the training pattern, and choose 200 samples as the test pattern. It means there 4000 training samples and 400 test samples. The classification error of 1-NN classifier for this example is 0.0%.

Then we train the ASC with Algorithm 3.5. Fig. 5 gives the adjusted SOINN result, it is nearly the same as the results of first-layer of SOINN. It shows that adjusted SOINN can represent the topology of two classes well. Fig. 6 is the training results of ASC. ASC removed lots of prototypes generated by adjusted SOINN, such removed prototypes are not useful for the following classification. It shows that to realize the classification, ASC only need 6 prototypes (3 for one class, and 3 for another class). With these 6 prototypes, we classify the test set and get 0.0% classification error. Compared with 1-NN classifier, ASC uses $6/4000 = 0.15\%$ (compression ratio $r_c = 0.15\%$) prototypes of 1-NN and get the same classification error.

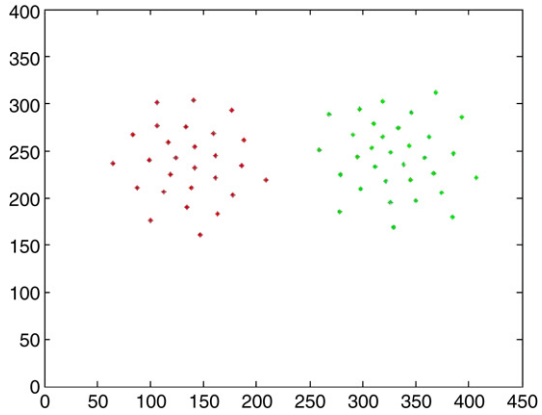

Fig. 5. Adjusted SOINN results of Fig. 4, it is nearly the same as the results of first-layer of SOINN.

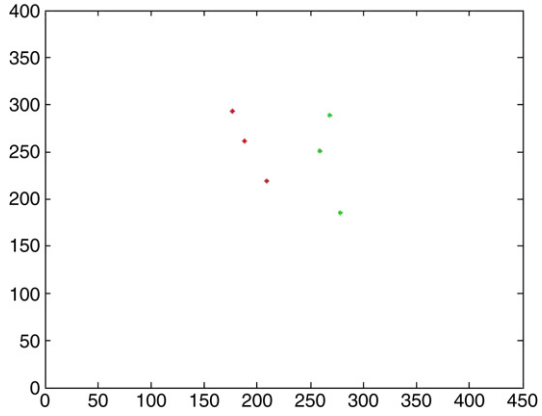

Fig. 6. ASC results of Fig. 4.

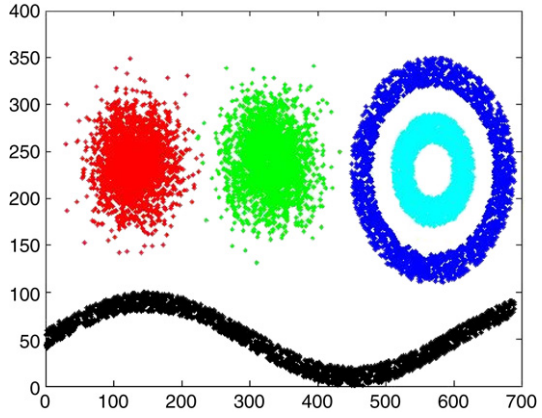

Fig. 7. Two Gaussian distribution, two concentric rings, and sinusoid curve, without overlap (5 classes).

(2) Experiment 2: five classes with different distribution and different shape, without overlap

In this experiment, we add three other classes to the classes in Experiment 1. The classes obey the distribution shown in Fig. 7. They are two Gaussian distribution, two concentric rings, and one sinusoid curve. We randomly took 2000 samples from every class as the training pattern, and randomly took 200 samples from every class as the test pattern, i.e. there are 10,000 samples in training set and 1000 samples in test set. The classification error of 1-NN is 0.0%.

Fig. 8 is the adjusted SOINN results. It represents the topology structure of input data very well. Fig. 9 is the ASC results. For the two Gaussian distribution classes, the number of prototypes becomes larger than in Experiment 1, it is because there are other classes outside the two classes, and to build decision boundary between the Gaussian distribution classes and other classes, it needs more prototypes. For the concentric rings and sinusoid, ASC also automatically determined the number of prototypes needed to decide the decision boundary. Fig. 9 also shows that for different classes, ASC allocates different number of prototypes for the reason that different classes obey different distribution or have different shape and size. The classification error of ASC is 0.0% and it needs 69 prototypes, i.e., ASC uses 69/10,000 = 0.69% (r_c = 0.69%) prototypes of 1-NN and get the same classification error.

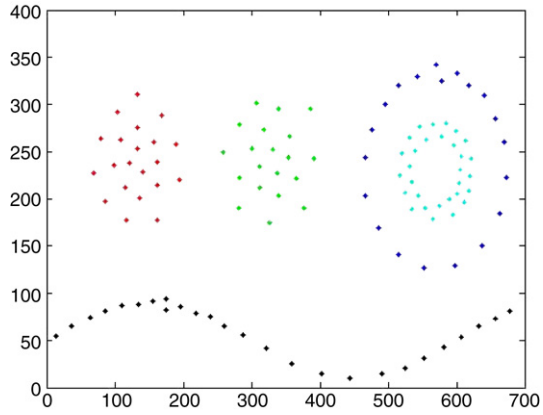

Fig. 8. Adjusted SOINN results of Fig. 7, it is nearly the same as the results of first-layer of SOINN.

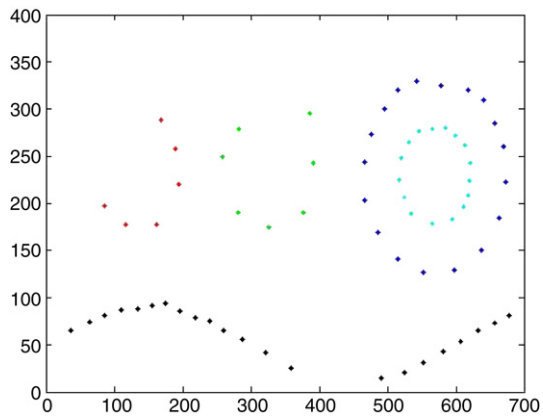

Fig. 9. ASC results of Fig. 7.

(3) Experiment 3: five classes with different distribution and different shape, with overlap

In this experiment, we adjust the classes in Experiment 2 by move two Gaussian distribution classes together to form overlapped classes. Fig. 10 is the distribution. We randomly took 2000 samples from every class as the training pattern, and randomly took 200 samples from every class as the test pattern. The classification error of 1-NN is 2.7%.

Fig. 11 is the adjusted SOINN results, and there are overlap prototypes between two Gaussian distribution classes. Fig. 12 is the ASC results. It shows that ASC separated the overlapped area. The classification error of ASC is 2.0% and it needs 86 prototypes, i.e. ASC uses 86/10,000 = 0.86% (r_c = 0.86%) prototypes of 1-NN and gets lower classification error for the overlapped classes.

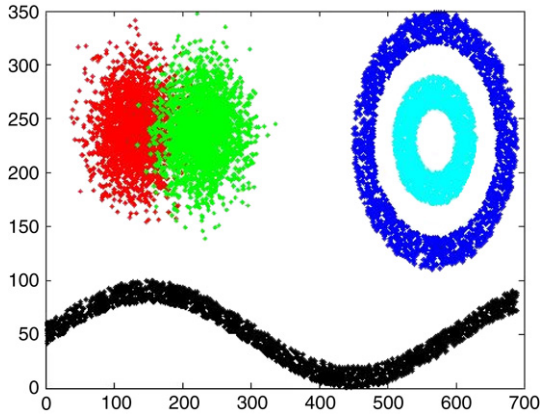

Fig. 10. Two Gaussian distribution with overlap, two concentric rings, and sinusoid curve (5 classes).

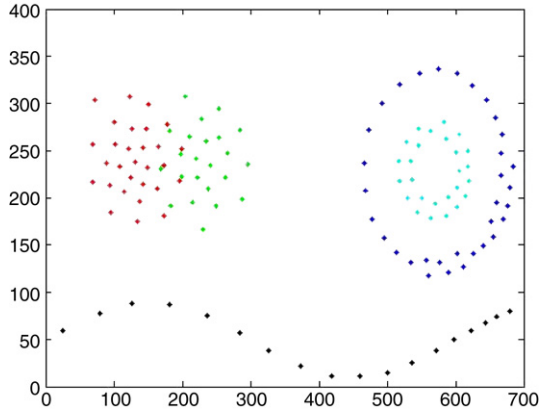

Fig. 11. Adjusted SOINN results of Fig. 10, it is nearly the same as the results of first-layer of SOINN.

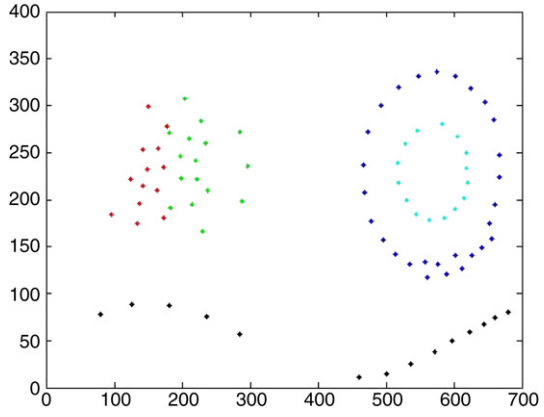

Fig. 12. ASC results of Fig. 10.

(4) Experiment 4: five classes with different distribution and different shape, with overlap and 10% noise

In this experiment, we add 10% noise to the classes in Experiment 3. The noise is randomly distributed on the whole feature space and we randomly label such noise samples with class names. Fig. 13 is the distribution, it shows that some samples belong to one class, but they have different class labels. We randomly took 2000 samples from every class as the training pattern, but for the test pattern we use the same test set as in Experiment 3. It means that in the training samples there are noise samples, some samples belong to one class but labeled as different classes. We use the test set without noise to test how the noise in the training data influence the recognition performance. The classification error of 1-NN is 5.9%.

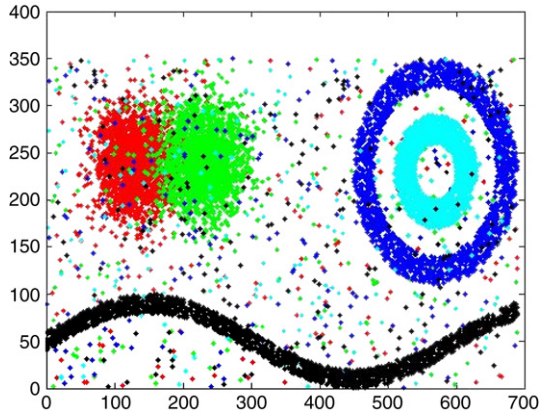

Fig. 13. Two Gaussian distribution with overlap, two concentric rings, and sinusoid curve; with 10% noise (5 classes).

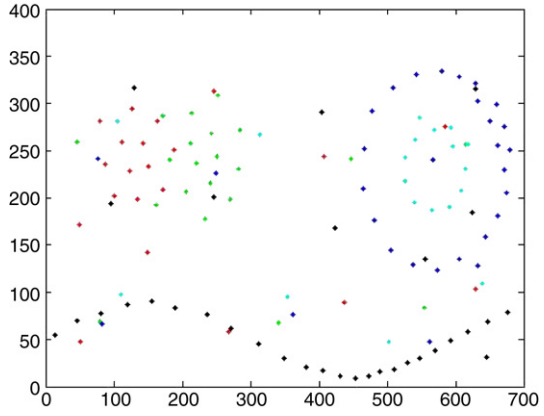

Fig. 14. Adjusted SOINN results of Fig. 13, it is nearly the same as the results of first-layer of SOINN.

Fig. 14 is the adjusted SOINN results. Even adjusted SOINN considered some noise delete function, there are still some prototypes caused by noise remain undeleted. Fig. 15 is the ASC results, the noise-reduction process deleted those noise prototypes, and the center-cleaning process removed the prototypes that are not useful for classification. The classification error of ASC is 2.2% and it needs 87 prototypes, i.e., ASC uses 87/10,000 = 0.87% (r_c = 0.87%) prototypes of 1-NN and gets much lower classification error. Compare Experiment 4 with Experiment 3, we can conclude that ASC is robust to noise for the reason that even there are 10% noise, ASC can get nearly the same classification error as the no noise training data with nearly the same compression ratio.

(5) Experiment 5: two spiral problem, with 2 classes

The spiral problem is a classic example of non-linear data (available from the Carnegie Mellon AI repository). If a total of N data points are to be generated, then the two dimension spiral shape data change as follows, for $1 <= i <= N$:

$$\text{angle} = (i \times \pi)/(16 \times \text{density}) \tag{15}$$

$$\text{radius} = \text{maxRadius} \times ((104 \times \text{density}) - i)/(104 \times \text{density}) \tag{16}$$

$$x = \text{radius} \times \cos(\text{angle}) + \text{offset} \tag{17}$$

$$y = \text{radius} \times \sin(\text{angle}) + \text{offset}. \tag{18}$$

Here x and y are the spiral data points, and $\pi = 3.14$. The two spirals are governed by three parameters: density φ,

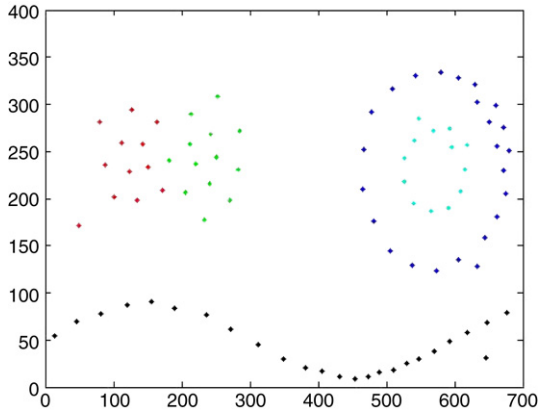

Fig. 15. ASC results of Fig. 13.

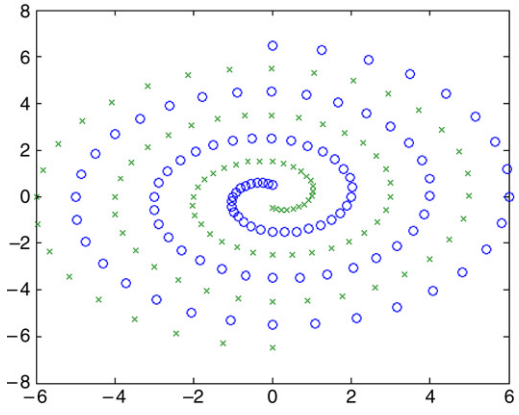

Fig. 16. Original data set of two spiral problem.

maximum radius ρ, and offset δ. The density variable defines the total number of points generated within an envelope defined by the radius. By different spiral parameters, it is possible to generate different spirals with varying radius and length. Fig. 16 was proposed with $\varphi = 1$, $\rho = 6.5$, and $\delta = 0$. Data belonging to two different classes lie on these two different spirals (represented as a sequence of 'o' and 'x' in Fig. 16).

The training set is represented by the vector $\{x, y\}$. The test sets may be thought of as noisy spirals, i.e. training set plus a uniform level of noise. For the two spiral problem, three test sets are possible: $\{x, y + \delta\}$, $\{x + \delta, y\}$, and $\{x + \delta, y + \delta\}$. It is possible to generate several different test sets of noisy spirals with varying δ.

In this experiment, to generate training data, we set $\rho = 6.5$, $\varphi = 10$, and $\delta = 0$. It generates 1940 samples for training. To generate testing data, we set $\rho = 6.5$, $\varphi = 1$, and $\delta = 0.1, 0.2, 0.3$. It generates 1746 samples for testing. During the training, we set the parameters as following: $a_d = 20$, $\lambda = 20$, $k = 1$. Fig. 17 is the training results of ASC. There are 187 nodes needed to represent the typical prototypes of classes. 79 prototypes are needed for class marked with 'o', and 108 prototypes are needed for class 'x'. The compression ratio is $187/1940 = 9.6\%$. The classification error are listed in Table 1.

4.2. Real world data: Character recognition

In this experiment, we use the Optical Recognition of Handwritten Digits database (optdigits) (Merz & Murphy, 1996) to test ASC. In this database, there are 10 classes (handwritten digits) from a total of 43 people, 30 contributed to the training set and different

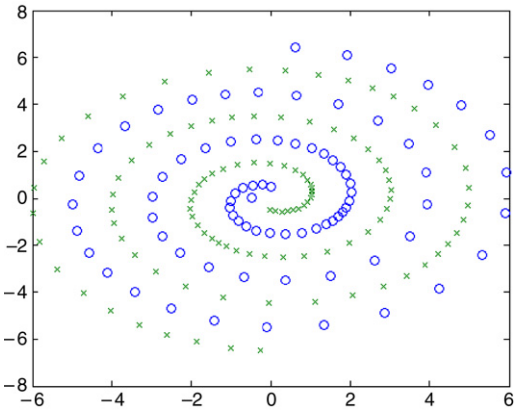

Fig. 17. Training results of ASC.

Table 1
Classification results of two spiral problem

δ	Classification error on $(x + \delta, y)$ (%)	Classification error on $(x, y + \delta)$ (%)	Classification error on $(x+\delta, y+\delta)$ (%)	Classification error on average (%)
0.1	1.5	2.6	2.5	2.2
0.2	2.3	3.2	2.9	2.8
0.3	2.4	3.5	3.2	3.0

Table 2
Number of samples in optdigits database

Class	No. of training samples	No. of test samples
0	376	178
1	389	182
2	380	177
3	389	183
4	387	181
5	376	182
6	377	181
7	387	179
8	380	174
9	382	180
Total	3823	1797

13 to the test set. The number of samples in the training set and testing set of every class are listed in Table 2, and there are a total of 3823 samples in the training set, and a total of 1797 samples in the test set. The dimension of the samples is 64. With traditional 1-Nearest Neighbor method (NNC, nearest neighbor classifier), using the 3823 training samples as the prototype vectors, we classify the test samples to different classes using Euclidean distance as the metric. The classification error is 2.0%. It spends 7 s and 919 ms (totally 7919 ms) to realize the classification of test samples.

As we pointed out in Section 3.3, for the noise-reduction part of ASC, the parameter k can be determined by cross-validation. We use 10-fold cross-validation to tune the parameter k in this experiment, and we get $k = 1$ as the tuning result.

During the training of ASC, we test 3 different parameter sets for (a_d, λ): (1) (50, 50), (2) (25, 25), and (3) (10, 10). We do 10 times training for every parameter set, and take the average results as the last results. Table 3 lists the number of prototypes of every class for different parameter sets. Using such prototypes, we classify the test samples to different classes and give the recognition results for different parameter sets in Table 4. From Table 3 we know that, for different classes, the needed number of prototypes is different, and larger $\{a_d, \lambda\}$ will lead to more prototypes; Table 4 shows that, (1) with the classification error 2.3% (Column II of Table 4), ASC speed up the traditional Nearest Neighbor method more than 10

Table 3
Number of prototypes for different classes of optdigits with different parameter sets (a_d, λ), displayed with average of 10 times training and standard deviation

Class	No. of nodes for different sets of (a_d, λ)		
	(50, 50)	(25, 25)	(10, 10)
0	32 ± 4	20 ± 3	11 ± 3
1	40 ± 4	31 ± 4	11 ± 1
2	38 ± 3	27 ± 2	12 ± 5
3	41 ± 5	24 ± 3	12 ± 3
4	35 ± 3	25 ± 2	9 ± 3
5	39 ± 5	28 ± 5	10 ± 3
6	34 ± 3	25 ± 2	11 ± 3
7	33 ± 3	24 ± 2	12 ± 2
8	39 ± 5	25 ± 4	12 ± 3
9	45 ± 5	30 ± 3	12 ± 2
Total number	377 ± 12	258 ± 7	112 ± 7

For different classes, the needed number of prototypes is different.

Table 4
Recognition performance of ASC classifier for optdigits with different parameter sets (a_d, λ), displayed with average of 10 times training and standard deviation

	Parameter set of $\{a_d, \lambda\}$		
	(50, 50)	(25, 25)	(10, 10)
Classification error (%)	2.3 ± 0.2	2.6 ± 0.2	3.0 ± 0.2
No. of prototypes	377 ± 12	258 ± 7	112 ± 7
Time needed (ms)	781	534	232
Speed up ratio (r_s(NNC, ASC))	10.14	14.82	34.13
Compression ratio (%)	9.9 ± 0.3	6.8 ± 0.2	2.9 ± 0.2

times (speed up ratio is 10.14) with less than 10% memory space ($r_c = 9.9\%$); (2) if we want to speed up the classification process much faster with much lower memory space, the classification error will be increased (Column II – Column IV of Table 4), and the parameter sets (a_d, λ) can be used to control the recognition performance.

For your reference, we also compared ASC with support vector machine (SVM) using the Optdigits dataset. We use the well-known LibSVM (http://www.csie.ntu.edu.tw/Ecjlin/libsvm/) to test the Optdigits dataset (with a Radius Basis Function as the kernel function); it uses 1197 support vectors to obtain a 3.4% classification error. With Column IV of Table 4, we know that ASC can get lower classification error with only 1/10 memory of LibSVM (classification error is 3.0%, number of prototypes is 112). Passerini et al. calculated the recognition ratio of multiclass SVM with different kernel functions (Passerini, Pontil, & Frasconi, 2002): for One-vs.-All SVM methods, the lowest classification error of Optdigits is 2.8%; for All-pairs SVM methods, the lowest classification error of Optdigits is 2.6%. Table 4 shows that the ASC can achieve equivalent or better recognition results than SVM.

4.3. Compare with other prototype-based methods

Veenman and Reinders compare the Nearest Subclass Classifier NSC(σ) (Veenman & Reinders, 2005) with some other classifiers include the k-Means Classifier KMC(M) (Hastie et al., 2001), the k-Nearest Neighbors Classifier NNC(k) (Cover & Hart, 1967), the k-Edited Neighbors Classifier ENC(k) (Wilson, 1972), the Multiscale Data Condensation algorithm MDC(k) (Mitra, Murthy, & Pal, 2002), the Bootstrap technique BTS(M) with $T = 100$ and $k = 3$ (Bezdek & Kuncheva, 2001), and Learning Vector Quantization LVQ(M) with $\alpha = 0.3$, $\eta = 0.8$, and $T = 100$ as in Bezdek and Kuncheva (2001). For the above methods, Veenman and Reinders optimize the parameter with tuning by cross-validation. They also compare NSC classifier with some methods with fixed parameter such as NNC(3), RT3(3), and TAB(α) with $\alpha = 0.05$, $T_t = 0.03N$, $T_i = 20$, and $T = 100$ as in Bezdek and Kuncheva (2001). They show that the tuning of parameters can greatly improve the recognition performance. They also prove that NSC classifier works better than a parameterless classifier MCS condensing algorithm (Dasarathy, 1994). In summary, Veenman and Reinders declare that NSC classifier works better than the above mentioned classifiers from the balance of recognition ratio and classification speed.

Table 5
Data sets used for comparison

Data set	Objects	Features	Classes
Iris	150	4	3
Breast cancer	683	9	2
Ionosphere	351	34	2
Glass	214	9	6
Liver disorders	345	6	2
Pima Indians	768	8	2
Wine	178	13	3

In this section, we compare the proposed ASC with some other prototype-based classifiers to evaluate the ASC, and we use the results in Table 2 of Veenman and Reinders (2005) to compare the proposed ASC with NSC, KMC, and LVQ. For the recognition performance of some other classifiers, please refer Veenman and Reinders (2005).

In Veenman and Reinders (2005), the authors use 9 datasets to evaluate NSC and compare it with other methods. These 9 datasets include one artificial dataset made by the authors, and a Sonar dataset with 30 features vectors. It is difficult for us to regenerate the artificial dataset and the 30-feature Sonar dataset (in general, Sonar dataset has 60 features vectors), thus we compare the proposed ASC with NSC and other classifiers with other 7 datasets. We list such datasets in Table 5. For the detail of such datasets, we can refer Merz and Murphy (1996).

As a cross-validation result for real-world data Optdigits in Section 4.2, the parameter k in the noise-reduction part is set as 1. Here, we also adopt $k = 1$ for all datasets in this section.

Other than k in ASC, there are two parameters that cannot be learned directly from the training data. We use a general 10-fold cross-validation to tune such parameters; the recognition results are shown in Table 6. Furthermore, for all seven real-world datasets, we also listed the parameter values obtained with tuning by cross-validation of all algorithms in Table 7, where we also show the achieved compression ratio of the algorithms. The classification speed is compared between ASC and other classifiers in Table 8, the speed up ratio of ASC to other classifiers are listed in Table 8.

With the first well-known Iris dataset, we compared ASC with other algorithms. This dataset consists of 150 objects that are subdivided in three classes. Each object contains four features. With the optimum parameter value for the ASC ($a_d = 6$, $\lambda = 6$), the number of prototypes per class is $M_1 = 2$, $M_2 = 2$, and $M_3 = 3$. The optimum result of NSC gave $M_1 = 2$, $M_2 = 3$, and $M_3 = 4$; KMC requires four prototypes per class; NNC uses all training data as prototypes; and LVQ needs 22 prototypes per class. The ASC obtains the lowest classification error among the algorithms. Comparing ASC with KMC and LVQ, the difference is that, for this dataset, the number of prototypes is not needed to be same for all classes, and one class can be modeled with far fewer prototypes than the other requires. Consequently, ASC can use fewer prototypes to achieve an even lower classification error. Comparing ASC with NSC, the difference is that: (1) Adjusted SOINN can represent the data distribution very well. (2) The center-cleaning process of ASC removed some prototypes that are not useful during the classification process. The two features enable ASC to obtain a lower classification error with a better compression ratio. For classification speed, ASC works 1.3 times faster than NSC, 1.7 times faster than KMC, 21 times faster than NNC, and 2.14 times faster than LVQ.

Table 6
Comparison results of ASC and other classifiers: Classification error in percentages, displayed with average of 10 times 10-fold cross-validation and standard deviation

Data set	ASC (a_d, λ)	NSC ($\sigma^2_{\max}$)	KMC (M)	NNC (k)	LVQ (M)
Iris	**2.6** ± 0.86	3.7 ± 0.4	3.8 ± 0.8	3.3 ± 0.6	3.9 ± 0.6
Breast cancer	**2.6** ± 0.38	**2.8** ± 0.2	4.1 ± 0.3	**3.0** ± 0.2	3.7 ± 0.4
Ionosphere	**9.6** ± 0.64	**8.1** ± 0.8	12.6 ± 0.6	13.9 ± 0.7	3.6 ± 0.8
Glass	**26.5** ± 1.6	29.8 ± 1.5	31.2 ± 1.1	27.7 ± 1.2	31.7 ± 2.0
Liver disorders	37.4 ± 0.83	37.1 ± 2.3	40.7 ± 2.3	**32.7** ± 1.6	33.7 ± 1.9
Pima Indians	28.0 ± 0.63	31.4 ± 1.6	31.3 ± 0.9	**25.3** ± 0.7	26.5 ± 0.9
Wine	**17.4** ± 1.55	24.7 ± 1.7	28.1 ± 1.9	26.2 ± 1.9	27.7 ± 1.5
Average	**17.7** ± 0.93	19.6 ± 1.2	21.7 ± 1.1	18.9 ± 0.99	20.1 ± 1.2

Bold face classification error is the best or near best classifier.

Table 7
Comparison results of ASC and other classifiers: Compression ratio in percentages, bold face compression ratio means the best or near best classifier

Data set	ASC (a_d^*, λ^*)	NSC ($\sigma^{2\,*}_{\max}$)	KMC (M^*)	NNC (k^*)	LVQ (M^*)
Iris	**5.2** (6, 6)	7.3 (0.25)	8.0 (4)	100 (14)	15 (22)
Breast cancer	1.4 (8, 8)	1.8 (35.0)	**0.29** (1)	100 (5)	5.9 (40)
Ionosphere	**3.4** (15, 15)	31 (1.25)	4.0 (7)	100 (2)	6.8 (24)
Glass	**13.7** (15, 15)	97 (0.005)	17 (6)	100 (1)	45 (97)
Liver disorders	**4.6** (6, 6)	**4.9** (600)	11 (19)	100 (14)	8.4 (29)
Pima Indians	**0.6** (6, 6)	1.7 (2600)	1.0 (4)	100 (17)	3.4 (26)
Wine	**3.2** (6, 6)	96 (4.0)	29 (17)	100 (1)	32 (57)
Average	**4.6**	34.2	10.0	100	16.6

The optimal parameter value tuned by cross-validation is shown in ().

Table 8
Comparison results of ASC and other classifiers: Classification speed

Data set	NSC r_s(NSC, ASC)	KMC r_s(KMC, ASC)	NNC r_s(NNC, ASC)	LVQ r_s(LVQ, ASC)
Iris	1.3	1.7	21	2.14
Breast cancer	1.28	0.21	71	4.2
Ionosphere	9.14	1.17	24.5	2
Glass	7.07	1.34	7.3	3.28
Liver disorders	1.06	2.39	21.7	1.82
Pima Indians	2.83	1.67	166.9	5.67
Wine	30	9.07	31.2	10.01
Average	7.53	2.5	49.1	4.16

Speed up ratio of ASC to other classifiers. If the speed up ratio is greater than 1, ASC works faster than the corresponding classifier.

The testing of other datasets also shows the efficiency of the proposed ASC. Table 6 shows that, for Iris, Breast cancer, Glass, and Wine datasets, ASC achieves the lowest classification error. For the Ionosphere and Pima Indian datasets, the classification error is near the best classifier. The last row in Table 6 shows the average of the classification error of all datasets and summarizes the performance of all algorithms. The data in that row shows that, compared with other prototype-based classifiers, ASC achieves the lowest classification error. Table 7 is the compression ratio. For all datasets, the ASC has the best compression ratio or nearly the best compression ratio. On average, ASC achieves the best compression ratio, which means that ASC needs less memory space and processing time than other classifiers. Table 8 compares the classification speed of ASC and other classifiers. In this table, if speed up ratio is greater than 1, ASC is faster than the corresponding method. For Breast cancer, KMC works a little faster then ASC, but for all other situation, ASC is the fastest classifier.

Tables 6–8 also show that, for some prototype-based classifier such as NNC and NSC, they can get low classification error, but they are time consuming. For some fast classifiers such as KMC and LVQ, the classification error is high. Compared with such methods, the proposed ASC achieves the lowest classification error with the fastest classification speed. Compared with the recently published NSC method, on average, ASC earns an 17.7% classification error with a 4.6% compression ratio; NSC needs a 34.2% compression ratio to achieve a 19.6% classification error; also, ASC can realize classification nearly 7 times faster than NSC. Results show that ASC works much more efficiently than NSC.

We must mention that, for several datasets (for example, Liver disorders, Pima Indians), the proposed ASC performs with a higher classification error than NNC, which is a well-established classifier in many domains. For NNC(k), Table 6 shows the tuning results of parameter k by cross-validation. k is 14 for the Liver Disorder dataset, and 17 for Pima Indians. Consequently, the classes in the Liver Disorder dataset (and the Pima Indian dataset) are heavily overlapped, i.e. the dataset has high density in the overlapped area. Fundamentally, ASC is a lossy classification method. During the processing of the dataset with very high-density overlapped areas, it is possible for ASC to lose some useful information for classification, and it is the reason that ASC performs with a higher classification error than NNC. However, the NNC(k) requires all training data to classify new objects, which is computationally expensive in terms of both time and storage. The computation loading is unbearable if it is necessary to use large k for NNC(k). Compare to NNC and other algorithms with these two datasets, ASC achieves the best compression ratio, the fastest classification speed, and the classification error is also comparable.

Finally, when storage and classification speed issues are considered, even ASC cannot yield the lowest classification error for some datasets; it gets the smallest average prototype set size (thus is needs smallest storage and realizes fastest classification) with the lowest average classification error compared to other prototype-based classifiers.

5. Conclusion

In this paper, we have proposed a new prototype-based classifier, based on adjusted self-organizing incremental neural

network (SOINN). We call this method the adjusted SOINN classifier (ASC). Using an adaptive similarity threshold, the system can grow incrementally and accommodate input patterns of incremental data distribution. By deleting the within-class insertion, the system requires fewer parameters than SOINN. The ASC can reduce the prototypes caused by noise and make it robust to noise and possible to achieve a low classification error. The deletion of unnecessary prototypes during the classification process makes ASC much faster than some other classifiers. In the experiment, ASC is compared with some other classifiers in terms of the classification error, compression ratio, and speed up ratio. ASC achieves the best performance and shows that it is a very efficient method.

We must mention that some other problems remain unresolved. For example, three parameters must be determined by the user or by some tuning algorithm: a_d and λ for adjusted SOINN process, and k for noise-reduction process. The difficulty of automatically determining such parameters is based on the fact that, for different tasks, the optimal choice of such parameters will differ. It is therefore difficult to give a standard for such parameters in all tasks. We mentioned incremental learning as a feature of ASC, but we cannot train the ASC with real-time data because the k-means process and center-cleaning process must store all training data. We will try to design an online incremental learning prototype-based classifier in the future.

Acknowledgment

This work was supported in part by the China NSF grant (#60573157, #60723003, and #60775046).

References

Bezdek, J. C., & Kuncheva, L. I. (2001). Nearest prototype classifier design: An experimental study. *International Journal of Intelligent Systems*, *16*, 1445–1473.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *IT-13*(1), 21–27.

Dasarathy, B. V. (1991). *Nearest neighbor (NN) norms: NN pattern classification techniques*. Los Alamitos, CA: IEEE CS Press.

Dasarathy, B. V. (1994). Minimal consistent set (MCS) identification for optimal nearest neighbor decision systems design. *IEEE Transactions on Systems, Man, and Cybernetics*, *24*(3), 511–517.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Kohonen, T. (1990). Improved versions of learning vector quantization. In *Proc. int'l joint conf. neural networks*: *Vol. 1* (pp. 545–550).

Merz, C., & Murphy, M. (1996). UCI repository of machine learning databases. Irvine, CA. University of California Department of Information. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Mitra, P., Murthy, C. A., & Pal, S. K. (2002). Density-based multiscale data condensation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(6), 734–747.

Passerini, A., Pontil, M., & Frasconi, P. (2002). From margins to probabilities in multiclass learning problems. In *Proceedings of the 15th European conference on artificial intelligence*.

Shen, F., & Hasegawa, O. (2006). An incremental network for on-line unsupervised classification and topology learning. *Neural Networks*, *19*, 90–106.

Shen, F., & Hasegawa, O. (2007). An enhanced self-organizing incremental neural network for online unsupervised learning. *Neural Networks*, *20*, 893–903.

Shen, F., & Hasegawa, O. (2005). An on-line learning mechanism for unsupervised classification and topology representation. In *IEEE computer society international conference on computer vision and pattern recognition*.

Veenman, C. J., & Reinders, M. J. T. (2005). The nearest subclass classifier: A compromise between the nearest mean and nearest neighbor classifier. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(9), 1417–1429.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, *2*(3), 408–421.

Wilson, D. R., & Martinez, T. R. (2000). Reduction techniques for instance-based learning algorithms. *Machine Learning*, *38*, 257–286.