Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 20 (2007) 819–831
Neural Networks
www.elsevier.com/locate/neunet

Partial distortion entropy maximization for online data clustering

Hiroyuki Takizawa[a,*], Hiroaki Kobayashi[b]

[a] Graduate School of Information Sciences, Tohoku University, 6-3 Aramaki-aza-aoba, Aoba, Sendai 980-8578, Japan
[b] Information Synergy Center, Tohoku University, 6-3 Aramaki-aza-aoba, Aoba, Sendai 980-8578, Japan

Received 8 February 2006; accepted 20 April 2007

Abstract

Competitive learning neural networks are regarded as a powerful tool for online data clustering to represent a non-stationary probability distribution with a fixed number of weight vectors. One difficulty in practical applications of competitive learning neural networks to online data clustering is that most of them require heuristically-predetermined threshold parameters for balancing a trade-off between convergence accuracy, i.e. error minimization performance, and speed of adaptation to the changes in source statistics. Although adaptation acceleration is achievable by relocating a "useless" node so that it becomes useful, excessive relocation often disturbs error minimization. Hence, both of the adaptation speed and the error minimization performance sensitively depend on threshold parameters to determine whether a node should be relocated or not. In general, it is difficult to know adequate threshold parameters *a priori*. This paper proposes a novel criterion for decision making of node relocation without heuristically predetermined thresholds. According to the proposed criterion, a node is relocated only if the relocation task improves *partial distortion entropy*, which is an online optimality metric reliable from the viewpoint of error minimization. Hence, node relocation is carried out without disturbing error minimization. As a result, both quick adaptation and error minimization are simultaneously accomplished without any carefully predefined parameters. Experimental results clarify the validity of the proposed criterion. Competitive learning with the criterion is clearly superior to other representative algorithms in terms of both quick adaptation and error minimization performance.

Keywords: Competitive learning; Partial distortion theorem; Non-stationary probability distributions; Partial distortion entropy maximization

1. Introduction

Unsupervised competitive learning has been developed to estimate unknown probability distributions of input vectors (Kohonen, 1989, 1995). In competitive learning, a distribution of input vectors is represented with a fixed number of weight vectors, each corresponds to a node of a competitive learning neural network (CLNN). CLNNs are often regarded as a powerful tool for data clustering that is to group a huge number of vectors into some classes (clusters) (Fayyad, Haussler, & Stolorz, 1996; Jain, Murty, & Flynn, 1999).

One advantageous feature of competitive learning for data clustering is that it is innately applicable to online clustering of time-series vectors drawn from a nonstationary probability distribution. A CLNN can be seen as an online version of the k-means data clustering algorithm (MacQueen, 1967). A CLNN iterates learning steps to compensate weight vectors, so that the mean squared error for the current probability distribution decreases. However, the basic competitive learning algorithm needs a lot of learning steps to minimize the mean squared error. Therefore, its computational cost also becomes expensive as input vectors and/or weight vectors increase, and it might fail in adaptation to drastic changes in the non-stationary probability distribution.

To minimize the error with fewer learning steps, many acceleration schemes for CLNNs have been proposed so far. In general, such CLNNs are particularly devised to achieve quick adaption to the changes in probability distributions. For drastically updating a node, hence, many CLNNs employ addition/duplication/splitting of a *useful* node based on their own criteria (Chen, Sheu, & Fang, 1994; Fowler, 1998; Gersho & Yano, 1991; Goldberg & Sun, 1988; Nakajima, Takizawa, Kobayashi, & Nakamura, 1998; Nishida, Kurogi, & Saeki, 2001; Paul, 1982; Shen, Zeng, & Liou, 2003; Yoshizawa, Doki, & Okuma, 1999). At the same time, a *useless* node is removed to keep the number of nodes constant. This operation can be regarded as relocating a useless node so that it becomes useful, referred to as *node relocation*.

* Corresponding author.
E-mail addresses: tacky@isc.tohoku.ac.jp (H. Takizawa), koba@isc.tohoku.ac.jp (H. Kobayashi).

0893-6080/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2007.04.029

An important research issue in node relocation is to define how to determine whether a node should be relocated or not. Despite the importance of node relocation, its performance is often determined by heuristically-predetermined parameters. Note that there is a trade-off between the error minimization performance and adaptation speed; excessive node relocation often disturbs error minimization against the current probability distribution instead of accelerating it.

This paper proposes a novel criterion for decision making of node relocation. The criterion does not require predetermined parameters; the threshold for node relocation is dynamically adjusted according to the optimality of weight vectors. The optimality is assessed based on the partial distortion theorem (Gersho, 1979; Gersho & Gray, 1992), which is reliable to achieve minimization of the mean squared error. Without any heuristically-predetermined thresholds, therefore, node relocation based on the criterion can allow weight vectors to quickly adapt to the changes in a nonstationary distribution, while efficiently compensating a weight vector so as to minimize the mean squared error. In addition, this paper presents a simple competitive learning algorithm with the proposed criterion.

The remainder of this paper is organized as follows. Section 2 describes online clustering discussed in this paper. In Section 2, a useful metric to dynamically estimate the optimality of weight vectors, referred to as *partial distortion entropy*, is also described, briefly reviewing the basic theory of an optimal quantizer for a stationary signal source. Section 3 proposes a criterion for node relocation based on the optimality metric. A competitive learning algorithm based on the criterion is also presented. In Section 4, the validity of the proposed criterion is verified through experiments. Section 5 presents concluding remarks and our future work.

2. Online data clustering

2.1. Preliminary definitions

Let $\boldsymbol{x}(t)$ ($\boldsymbol{x}(t) \in \mathcal{R}^k$) be an input vector drawn from non-stationary probability density function, $P_t(\boldsymbol{x})$ at time t, and $\boldsymbol{Y}(t)$ be a set of weight vectors, i.e. $\boldsymbol{Y}(t) = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$. A CLNN maps any vector within $\mathcal{R}^k$ into one of a finite set of nodes, each corresponds to its own weight vector:

$$Q_t : \mathcal{R}^k \to \boldsymbol{Y}(t). \tag{1}$$

Hence every input vector $\boldsymbol{x}$ is mapped into the node of the nearest weight vector, $\boldsymbol{y}_w(t)$:

$$Q_t(\boldsymbol{x}(t)) = \boldsymbol{y}_w(t), \tag{2}$$

where w is the index of the node corresponding to the nearest weight vector, called a *winner*:

$$w = \arg\min_i \|\boldsymbol{x}(t) - \boldsymbol{y}_i(t)\|. \tag{3}$$

The mean squared error is defined by

$$D(t) = \sum_{i=1}^{N} \int_{S_i} \|\boldsymbol{x}(t) - \boldsymbol{y}_i(t)\|^2 \cdot P_t(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{4}$$

where N is the number of nodes and S_i is the Voronoi region of $\boldsymbol{y}_i$:

$$S_i = \{\boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{y}_i\| \le \|\boldsymbol{x} - \boldsymbol{y}_j\|;\ j = 1, 2, \ldots N, i \ne j\}, \tag{5}$$

and

$$\bigcup_j S_j = \mathcal{R}^k. \tag{6}$$

Suppose that, as commonly assumed in many literatures (e.g. Gersho and Gray (1992), Goldberg and Sun (1988)), a nonstationary probability distribution of input vectors can be approximated by a sequence of locally-stationary, ergodic processes. Hence, the series of input vectors over time, $X = \{\boldsymbol{x}(t)\}$, is divided into disjoint and exhaustive *frames* (Gersho & Gray, 1992), each of which is regarded as a finite/infinite subset of input vectors that arise from an unknown stationary probability distribution. Thus, the probability density function of the distribution, $P_t(\boldsymbol{x})$, is time-invariant within one frame and changes only when transiting to the next frame.

Let $P_\tau(\boldsymbol{x})$ be the short-term stationary probability density function within the τ-th frame F_τ, i.e. $P_\tau(\boldsymbol{x}) = P_t(\boldsymbol{x} | t \in F_\tau)$. Then the probability density function of a nonstationary distribution can be expressed by a set of time-sharing stationary probability density functions, $\{P_\tau(\boldsymbol{x}) | \tau = 1, 2, \ldots\}$, where $F_\tau \cap F_{\tau'} = \emptyset$ for $\tau \ne \tau'$.

This paper addresses an online data clustering framework that compensates a weight vector to minimize the error for the current frame defined by

$$D_\tau = \sum_{i=1}^{N} \int_{S_i} \|\boldsymbol{x} - \boldsymbol{y}_i\|^2 \cdot P_\tau(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{7}$$

Under the assumption of a nonstationary probability distribution consisting of stationary frames, we can concentrate on error minimization within a frame, and therefore adopt offline data clustering techniques to online one. The allowable length of a frame is constrained by the adaptation speed of online data clustering. The more rapidly a data clustering algorithm can adapt to changes, the shorter frame the algorithm can cluster. Since the length of each frame is unknown in advance, quick adaptation is mandatory for tracking a nonstationary probability distribution. In addition to error minimization performance and adaptation speed, the requirements for online data clustering are given as follows:

(1) Robustness to initial weight vectors: a set of weight vectors with the minimum error for $F_{\tau-1}$ may be a poor set of initial weight vectors for error minimization for F_τ. Accordingly, we must employ acceleration techniques whose features do not depend on initial states of weight vectors.
(2) No time dependency: many nonadaptive competitive learning algorithms employ time-dependent parameters, which monotonically increase/decrease with time, in order to achieve better error minimization performance by fine-tuning weight vectors (e.g, Kohonen (1989), Kohonen (1995), Nielsen (1990), Takizawa, Sano, Nakajima, Kobayashi, and Nakamura (2003) Ueda and Nakano (1994), Zhu and Po (1998)). However, CLNNs cannot use