Semi-supervised Self-organizing Feature Map for Gene Expression Data Classification

Moumita Roy, Anwesha Law, and Susmita Ghosh

Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, India

Abstract. In this article, a one-dimensional self-organizing feature map (SOFM) neural network integrated with semi-supervised learning is used to predict the class label of gene expression data under the scarcity of the labeled patterns. Iterative learning of the semi-supervised SOFM network is carried out using a few labeled patterns along with some selected unlabeled patterns. The unlabeled patterns, for which the maximum target value is greater than a threshold, are selected as the confident ones. Results are found to be encouraging.

Keywords: Semi-supervised learning, gene expression data, self-organizing feature map.

1 Introduction

Microarray technology is an important tool used to monitor expression levels of genes of a given organism. A microarray is a glass slide onto which DNA molecules are fixed at specific locations called spots or features. A microarray may contain thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a particular gene. Gene expression data is mainly used to diagnose (predict class labels) a patient depending on his medical condition [1]. One of its major applications is precise and early diagnosis of cancer malignancies [5] which is quite difficult but very crucial for successful treatment. Gene expression technology integrated with some accurate statistical methods could play a better role for cancer diagnosis than traditional clinical factors. Given microarray experiments and some information about the outcome of the disease from some former patients, the task of class prediction (malignancies present or not) is to learn the relation between the gene expression levels and their corresponding outcome. Afterwards, a diagnosis of disease development for new patients can be carried out using their gene expression profiles. There are mainly two approaches used for class prediction of gene expression data: supervised and unsupervised. Between these, applicability of supervised method [5] is lesser due to insufficient available knowledge; whereas, prior information is not necessary for unsupervised methods and hence they are widely applied to find groups of co-regulated genes, categorize them, and predict information about an independent sample [11]. A number of studies have been done in this aspect [3,10].

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 688–694, 2013.
© Springer-Verlag Berlin Heidelberg 2013

Microarray experiments generate large datasets with expression values for thousands of genes (features), but usually for a very few samples (i.e., patients). The labeling of the samples require expert knowledge, which, in turn, gives rise to a situation of (labeled) data scarcity. Due to insufficiency of training samples, supervised methods cannot be developed. Also, if an unsupervised method is used, knowledge of labeled samples, though very little, will remain unused. In this situation, semi-supervised approach [4,6] could be opted instead of an unsupervised or a supervised one. Semi-supervision uses a small amount of labeled patterns with abundant unlabeled ones for learning, and integrates the merits of both supervised and unsupervised strategies to make full utilization of collected patterns [4]. Usefulness of semi-supervised approaches is already explored to predict the class label of gene expression data [9].

Artificial neural networks (ANNs) [7], one of the powerful soft-computing tools, have widespread applications in pattern recognition and image processing [2]. ANNs have the most appreciable quality of learning with or without supervision. Neural networks use nonparametric statistical approaches for pattern classification where no prior knowledge for input pattern distribution is required. Due to massive parallel nature of neural networks, results can be obtained in less time. Applications of various types of neural networks, e.g., self-organizing feature maps (SOFM) [8], multilayer perceptron [7] are already made in the area of class prediction [10,12] of gene expression data. But, as to the knowledge of the authors, neural network based semi-supervised approaches are not found for gene expression data classification. This motivated us to pursue the present study using neural networks in semi-supervised learning platform to improve the performance of class prediction of gene expression data.

In the proposed work, SOFM neural network [6,7] embedded with semi-supervised learning (termed as, SS-SOFM) is used for class prediction of microarray data using a few labeled data along with the selected (confident) unlabeled patterns. To check the effectiveness of the proposed approach, experiments are carried out on five preprocessed gene expression datasets [5]. Comparative analysis has been carried out between the proposed semi-supervised strategy and the corresponding unsupervised method using SOFM.

2 Kohonen's Self-organizing Feature Map

Self-organizing feature map (SOFM) network [7,8] learns iteratively and generates topological map of input patterns gradually. The network architecture consists of two layers: input and output. The number of neurons in the input layer is equal to the number of features of the input pattern. Here, one dimensional output layer is used and the number of neurons in the output layer is equal to the number of classes. Each neuron in the output layer is connected to all the neurons in the input layer. The connection weights are initialized randomly between $[0, 1]$. Training of the SOFM is continued epoch by epoch till convergence by following three consecutive steps: competition, co-operation and weight updating.

Let, the p^{th} input pattern be denoted by $\overrightarrow{X_p} = [x_{p,1}, x_{p,2}, ..., x_{p,y}]$, where y is the number of features of the input pattern. The y-dimensional weight vector of the i^{th} neuron in the output layer connected with all the neurons of the input layer at epoch ep is represented by $\overrightarrow{W_i^{ep}} = [w_{i,1}^{ep}, w_{i,2}^{ep}, ..., w_{i,y}^{ep}]$ $(i = 1, 2, ..., z$, where z is the total number of output neurons).

In the first step, the winner neuron j at epoch ep in the output layer is selected as the best-matched neuron for the p^{th} input pattern where the similarity measure $(\overrightarrow{W_j^{ep}} \cdot \overrightarrow{X_p})$ is the maximum. Now, the winning neuron co-operates with its neighbors. Let $h_{i,j}^{ep}$ denote the influence of the winning neuron j on its topological neighbor i at epoch ep and $D_{i,j}$ denotes the lateral distance between the winning neuron j and its neighbor i. The topological neighborhood $h_{i,j}^{ep}$ is defined in a way such that it attains the maximum value at the winning neuron j for which $D_{i,j}$ is zero and decreases monotonically with increasing lateral distance $D_{i,j}$. Here, $h_{i,j}^{ep}$ is defined as follows:

$$h_{i,j}^{ep} = \exp\left(-\frac{{D_{i,j}}^2}{2(\sigma^{ep})^2}\right), \tag{1}$$

where σ^{ep} represents the radius of the topological neighborhood at epoch ep. The size of this radius decreases with increase in ep.

The synaptic weight is then updated using the following equation:

$$\overrightarrow{W}_i^{ep+1} = \overrightarrow{W}_i^{ep} + h_{i,j}^{ep}\ \eta^{ep}\ (\overrightarrow{X_p} - \overrightarrow{W}_i^{ep}), \tag{2}$$

where, η^{ep}, the learning rate in epoch ep, decreases with increase in ep.

Table 1. Algorithmic representation of the proposed work

Step 1:	Collect a few labeled patterns from each of the classes.
Step 2:	Initialize (randomly) the connection weights of SOFM.
Step 3:	Train the SOFM with the few collected labeled patterns until convergence.
Step 4:	For all the unlabeled patterns, predict the output for each of the classes by passing them through the (trained) SOFM.
Step 5:	For all the unlabeled patterns, estimate the target values using the output values of itself and its K nearest neighbours.
Step 6:	Select the confident unlabeled patterns, for which the maximum target value is greater than a threshold.
Step 7:	Train the SOFM with the same set of labeled patterns along with the set of selected confident unlabeled patterns.
Step 8:	Repeat Steps 4-7 until convergence. At convergence, goto Step 9.
Step 9:	Assign hard class labels to all the unlabeled patterns.

3 The Proposed Semi-supervised SOFM Algorithm

As mentioned in Section 1, in the present work semi-supervised SOFM neural network is used for class prediction of gene expression data where a few labeled

patterns are available and they are required for semi-supervised learning. For the labeled patterns, if it belongs to the l^{th} class, then the target values and output values for the said class are both assigned to 1 and for the rest of the classes these values are assigned to 0 and are kept fixed during training as well as during testing phases. The algorithmic representation of the proposed methodology is given in Table 1 and detailed description is presented in subsequent sections.

3.1 Training of the SOFM

In the proposed work, learning process of SOFM is modified to incorporate the label information of the patterns. During training, the patterns are fed to the network one by one. At epoch ep (in the training step st), for p^{th} pattern, the output value $s_{p,l}^{ep}$ in the l^{th} class (here, each of the classes is represented by a neuron in the output layer) is estimated by computing the similarity measure between $\overrightarrow{X_p}$ and $\overrightarrow{W_l^{ep}}$, and is taken as,

$$s_{p,l}^{ep} = \overrightarrow{X_p} \cdot \overrightarrow{W}_l^{ep} = \sum_{k=1}^{y} x_{p,k} \cdot w_{l,k}^{ep}. \tag{3}$$

As already mentioned in Section 1, in the present work, iterative learning of SOFM is carried out using the labeled as well as the selected unlabeled patterns. Unlike the traditional SOFM, here, an output neuron is treated as the winning neuron for a given input pattern $\overrightarrow{X_p}$ if the target value of the pattern in the corresponding class is the maximum. In this way, we have incorporated the label information during learning of the network. Thereafter, the weight vectors of the winning neuron as well as its neighboring neurons are updated using equation (2). Learning using labeled patterns is continued epoch by epoch until convergence. To check convergence, the sum of square error between the target value and the estimated output value at epoch ep, denoted as O^{ep}, is calculated using

$$O^{ep} = \sum_{p=0}^{N} \sum_{l=1}^{C} (t_{p,l}^{st} - s_{p,l}^{ep})^2, \tag{4}$$

where N is the total number of input patterns and $t_{p,l}^{st}$ is the target value of the p^{th} pattern for the l^{th} class at training step st. Here, C denotes the number of classes. The weight updating is performed until $(O^{ep} - O^{ep-1}) < \delta$, where δ is a very small positive quantity. After each updating, the components of the weight vector $\overrightarrow{W_l^{ep}}$ are normalized to lie in $[0, 1]$. After every epoch, the values of learning rate and the size of the topological neighborhood are decreased.
Since, in the first training step, we do not have defined target values for the unlabeled patterns, the connection weights of the SOFM network are updated using the labeled patterns only.

3.2 Estimation of Target Value for Unlabeled Patterns

After convergence, the unlabeled patterns are presented to the network and the output in all the classes for each of the unlabeled patterns are predicted using

equation (3). Then, soft class label for each unlabeled pattern is estimated by averaging the corresponding output values of its K nearest neighbors thereby incorporating the labeled information from neighbors. The estimated soft class label is used for assigning the target value of the unlabeled patterns.

3.3 Collection of Unlabeled Patterns and Iterative Training of the SOFM

As mentioned earlier, in the present work, learning of SOFM is carried out iteratively using the labeled patterns along with the selected unlabeled patterns. Here, an unlabeled pattern, for which its maximum target value is greater than a threshold, is selected as the confident one. The threshold is obtained by averaging the summation of maximum target values for all the unlabeled patterns.
Training of SOFM, estimation of the target values for all the unlabeled patterns, and collection of the (confident) unlabeled patterns for the next training step are carried out iteratively until convergence or the number of training steps exceeds a prespecified value. To check convergence, at the end of each training step, the sum of square error (ξ_{st}) between the predicted support values using the (trained) SOFM network and the updated target values using the estimated support values of K nearest neighbors for all the unlabeled patterns is computed. The training steps are performed until $(\xi_{st} - \xi_{st-1}) < \epsilon$, where ϵ is a very small positive quantity. After convergence, the hard class labels are assigned to the unlabeled patterns depending on their target values.

4 Experimental Details and Analysis of Results

4.1 Datasets Used

To evaluate the effectiveness of the proposed SOFM based semi-supervised method (denoted as, SS-SOFM), experiments are conducted on five gene expression datasets denoting different types of cancers. Preprocessed versions of the datasets are available in [5]. In the present work, an input pattern consists of gene expression levels of a particular patient. The feature values of the input patterns are normalized between 0 to 1. Detailed description of the data sets are given below:

A. Brain tumor dataset: This dataset contains microarray gene expression profiles of 42 patients having 5 different types of tumors of the central nervous system. It has 10 medulloblastomas, 10 malignant gliomas, 10 atypical teratoid/rhabdoid tumors (AT/RTs), 8 primitive neuro-ectodermal tumors (PNETs) and 4 human cerebella. The raw data consists of 5,597 genes.

B. Colon cancer dataset: This dataset consists of expression levels of 40 tumor and 22 normal colon tissues for 2000 human genes.

C. Leukemia dataset: Gene expression levels of 72 patients either suffering from acute lymphoblastic leukemia (ALL, 47 cases) or acute myeloid leukemia (AML, 25 cases) are present in this dataset. It has a total of 3,571 genes.

D. Lymphoma dataset: There are a total of 62 samples, and the expression of 4,026 genes. This dataset contains gene-expression levels of the 3 most prevalent adult lymphoid malignancies, i.e., 42 samples of diffuse large B-cell lymphoma (DLBCL), 9 observations of follicular lymphoma (FL), and 11 cases of chronic lymphocytic leukemia (CLL).

E. Prostate cancer dataset: It has expression of 6,033 genes each, from 102 samples, of which 52 are of prostate tumors and 50 are from non-tumor prostate samples.

4.2 Results and Analysis

To assess the effectiveness of the proposed approach, experiments are conducted on five gene expression datasets. Performance of the proposed technique is compared with that of the corresponding unsupervised one using one dimensional SOFM neural network. To execute the proposed semi-supervised algorithm, for all the data sets, a single pattern from each of the classes are chosen randomly as labeled patterns. To find out K number of nearest neighbors, the value of K has been set to 9 for all the datasets. Results corresponding to semi-supervised and unsupervised techniques are depicted in Table 2. The best percentage of accuracy value (denoted as, PA) obtained over 20 different simulations and its corresponding Rand and Jacard indices are used as the performance measuring criteria. The values of Rand and Jacard coefficients lie between [0,1] and for both the indices, higher value signifies better class prediction. From the results, it is seen that for all the cases the proposed methodology outperforms the corresponding unsupervised technique in terms of all the measuring indices used in our experiment. The results also strengthen the fact that the use of a very small number of labeled patterns increases the performance of the system significantly.

Table 2. Comparative results

Datasets	Techniques	#Training Patterns	PA	Rand	Jacard
Brain Tumor	Unsupervised	-	71.4286	0.8200	0.3673
	SS-SOFM	5	80.9524	0.8769	0.5810
Colon Cancer	Unsupervised	-	58.0645	0.5050	0.3514
	SS-SOFM	2	70.9677	0.5447	0.5348
Leukemia	Unsupervised	-	76.3889	0.6342	0.4771
	SS-SOFM	2	97.2222	0.9452	0.9051
Lymphoma	Unsupervised	-	70.9677	0.6790	0.4467
	SS-SOFM	3	88.7097	0.8297	0.6840
Prostate	Unsupervised	-	60.7843	0.4952	0.3352
	SS-SOFM	2	75.4902	0.6263	0.4548

5 Conclusion

In this paper, a semi-supervised self-organizing feature map neural network is designed for more accurate prediction of class information for gene expression data. It is achieved by improvising semi-supervision along with the self-organizing capability of Kohonen's network. The network is trained by using a small number of labeled patterns along with a large amount of unlabeled patterns. From the results, it has been found that the proposed algorithm has an edge over the corresponding unsupervised version. Like other semi-supervised methods, the proposed technique also computation intensive.

References

1. Bhattacharya, A., De, R.K.: Average correlation clustering algorithm (ACCA) for grouping of co-regulated genes with similar pattern of variation in their expression values. Journal of Biomedical Informatics 43(4), 560–568 (2010)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
3. Chandrasekhar, T., Thangavel, K., Elayaraja, E.: Gene expression data clustering using unsupervised methods. In: Proceedings of the 3rd International Conference on Advanced Computing (ICoAC), pp. 146–150 (2011)
4. Chapelle, O., Schölkopf, B., Zien, A.: Semi-supervised Learning. MIT Press (2006)
5. Dettling, M.: BagBoosting for tumor classification with gene expression data. Bioinformatics 20(18), 3583–3593 (2004)
6. Ghosh, S., Roy, M.: Modified self-organizing feature map neural network with semi-supervision for change detection in remotely sensed images. In: Kuznetsov, S.O., Mandal, D.P., Kundu, M.K., Pal, S.K. (eds.) PReMI 2011. LNCS, vol. 6744, pp. 98–103. Springer, Heidelberg (2011)
7. Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd edn. Prentice-Hall of India, New Delhi (2008)
8. Kohonen, T.: Self-Organizing Maps, 2nd edn. Springer, Berlin (1997)
9. Maraziotis, I.A.: A semi-supervised fuzzy clustering algorithm applied to gene expression data. Pattern Recognition 45(1), 637–648 (2012)
10. Nikkila, J., Törönen, P., Kaski, S., Venna, J., Castrén, E., Wong, G.: Analysis and visualization of gene expression data using Self-Organizing Maps. Neural Networks 15(8-9), 953–966 (2002)
11. Srinivas, V.R.: Bioinformatics: A Modern Approach, 2nd edn. Prentice-Hall of India Pvt. Ltd., New Delhi (2007)
12. Valentini, G.: Supervised gene expression data analysis using support vector machines and multi-layer perceptrons. In: Proceedings of the 6th International Conference on Knowledge-Based Intelligent Information & Engineering Systems (2002)