Multi-class Cancer Classification by Semi-supervised Ellipsoid ARTMAP with Gene Expression Data

Rui Xu[1], Georgios C. Anagnostopoulos[2], and Donald C. Wunsch II[1]
[1]Applied Computational Intelligence Laboratory, Dept. of Electrical and Computer Engineering
University of Missouri – Rolla, Rolla, MO 65409-0249 USA
[2]Dept. of Electrical & Computer Engineering, Florida Institute of Technology, Melbourne, FL 32901-6975 USA

Abstract—**To accurately identify the site of origin of a tumor is crucial to cancer diagnosis and treatment. With the emergence of DNA microarray technologies, constructing gene expression profiles for different cancer types has already become a promising means for cancer classification. In addition to binary classification, the discrimination of multiple tumor types is also important. Semi-supervised Ellipsoid ARTMAP (ssEAM) is a novel neural network architecture rooted in Adaptive Resonance Theory suitable for classification tasks. ssEAM can achieve fast, stable and finite learning and create hyper-ellipsoidal clusters inducing complex nonlinear decision boundaries. Here, we demonstrate the capability of ssEAM to discriminate multi-class cancer through analyzing two publicly available cancer datasets based on their gene expression profiles.**

Keywords—**Gene expression data, Semi-supervised Ellipsoid ARTMAP, Cancer classification.**

I. Introduction

With the emergence and rapid advancement of DNA microarray technologies [1, 2], cancer classification through identification of the corresponding gene expression profiles has already attracted numerous efforts from a wide variety of research communities. Cancer classification is important to the subsequent diagnosis and treatment. Without the correct identification of cancer types, it is rarely possible to provide useful therapies and achieve expected effects. Traditional classification methods are largely dependent on the morphological appearance of tumors, parameters derived from clinical observations, and other biochemical techniques. Their applications are limited by the existing uncertainties and their prediction accuracy needs further improvement [3]. DNA microarray technologies offer caner researchers a new method to investigate the pathologies of cancer from a molecular angle under a systematic framework, and further, to make more accurate prediction in prognosis and treatment.

In practice, it is more common to discriminate more than two types of cancers. Ramaswamy *et al*. divided the multi-class problem as a series of binary classification sub-problems through either one-versus-all or all-pairs approach and employed support vector machines, weighted voting, and *k*-nearest-neighbors methods to distinguish 14 different tumor types [4]. Khan *et al*. trained Perceptrons to categorize small round blue-cell tumors (SRBCTs) with 4 sub-classes [5]. Furthermore, Scherf *et al*. constructed a gene expression database to study the relationship between genes and drugs for 60 human cancer cell lines originating from 10 different tumors, which provides an important criterion for therapy selection and drug discovery [6].

Here, we use a new neural network architecture – semi-supervised Ellipsoid ARTMAP (ssEAM) [7], which is based on Adaptive Resonance Theory (ART) [8], to analyze publicly accessible datasets on cancer research. ssEAM is capable of learning associative maps between clusters of an input and an output space, and has the properties of fast, stable and finite learning. Also, ssEAM can create nonlinear boundaries by using hyper-ellipsoids to represent the generated categories. We demonstrate the potential of ssEAM, combined with a simple gene selection technique, in successfully addressing the challenge of analyzing and interpreting massive, multidimensional gene expression data with computational efficiency and satisfying results, which are comparable to or better than those obtained by other classifiers.

The paper is organized as follows. Section II presents a brief introduction to ssEAM and experimental methods. The results of experiments are presented and discussed in section III and section IV concludes the paper.

II. Methods

ssEAM came as an enhancement and generalization of Ellipsoid ART (EA) and Ellipsoid ARTMAP (EAM) [9], which, in turn, follow the same learning and functional principles of Fuzzy ART (FA) and Fuzzy ARTMAP (FAM) [10]. EAM employ EA categories for the task of data aggregation, whose geometric representations, which are called categories, are hyper-ellipsoids embedded in the feature space. A typical example of such a category representation, when the input space is 2-dimensional, is provided in Fig. 1, where it is shown that each category *j* is described by its center location m_j, its orientation d_j, and a Mahalanobis radius M_j [9]. The shaded area in the figure constitutes the representation region of category *j*. A category encodes whatever information the EAM classifier has learned about the presence of data and their associated

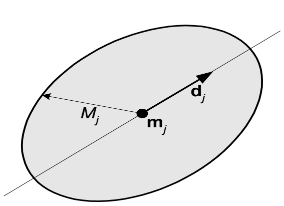

Fig. 1. Example of the geometric representation of an EAM category *j* when the feature space is 2-dimensional

class labels in the locality of its geometric representation. This information is encoded into the location and size of the hyper-ellipsoid. The latter feature is primarily controlled via the baseline vigilance $\bar{\rho} \in [0,1]$. Typically, small values of $\bar{\rho}$ produce categories of larger size, while values close to 1 produce the opposite effect. As a special case, when $\bar{\rho} = 1$, EAM consists solely of point categories (one for each training pattern) and implements the ordinary, Euclidian 1-Nearest Neighbor classification rule. A category's particular shape (eccentricity of its hyper-ellipsoid) is controlled via a network parameter $\mu \in (0,1]$; for $\mu = 1$ the geometric representations become hyper-spheres.

Fig. 2 illustrates the block diagram of an EAM network. EAM consists of two EA modules (ART$_a$ and ART$_b$) interconnected via an inter-ART module. The ART$_a$ module clusters patterns of the input domain and ART$_b$ the ones of the output domain. The information regarding the input-output associations is stored in the weights $\boldsymbol{w}_j^{ab}$ of the inter-ART module, while EA category descriptions are contained in the template vectors $\boldsymbol{w}_j$. These vectors are the top-down weights of F_2-layer nodes in each module.

The Semi-supervised EAM classifier extends the generalization capabilities of EAM by allowing the clustering into a single category of training patterns not necessarily belonging to the same class. This is being accomplished by augmenting EAM's prediction test (PT) in the following manner: a winning category J may be updated by a training pattern $\boldsymbol{x}$, even if the label of J is not equal to the class label of $\boldsymbol{x}$, as long as the following inequality holds:

$$w_{J,I(J)} \Big/ \left(1+\sum_{c=1}^{C} w_{J,c}\right) \geq 1-\varepsilon , \qquad (1)$$

where C denotes the number of distinct classes related to the classification problem at hand and the quantities $w_{j,c}$ contain the count of how many times category j was updated by a training pattern belonging to the c^{th} class. In other words, (1) ensures that the percentage of training patterns that are allowed to update category J and carry a class label different than the class label $I(J)$ (the label that was initially assigned to J, when it was created) cannot exceed 100ε %, where $\varepsilon \in [0,1]$ is the category prediction error tolerance parameter, which is specific only to ssEAM. For $\varepsilon = 1$ the modified PT will allow categories to be formed by clustering together training patterns regardless of their class labels in an unsupervised manner. In contrast, with $\varepsilon = 0$ the modified PT will allow clustering (into a single category) only of training patterns belonging to the same class, which makes the category formation process fully-supervised. Under these circumstances ssEAM becomes equivalent to EAM. For intermediate values of ε, the category formation process is performed in a semi-supervised fashion.

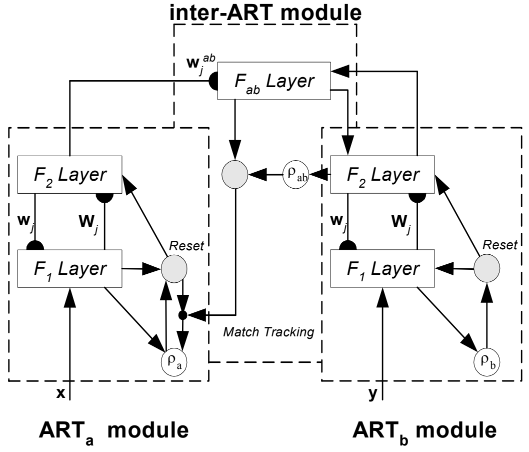

Figure 2: Ellipsoid ARTMAP block diagram

Due to its design, ssEAM has many attractive characteristics of learning, which are very desirable, for clustering or classification tasks. First, ssEAM is capable of both on-line and off-line learning. Using fast learning [7] in off-line mode, the network's training phase completes in a small number of steps. The computational cost during training is relatively low and it can cope with large amounts of multidimensional data, maintaining efficiency. Moreover, ssEAM is an exemplar-based model, that is, to accomplish the learning objective, during its training the architecture summarizes data via the use of exemplars. Due to its exemplar-based nature, responses of an ssEAM architecture to specific test data are easily explainable, which makes ssEAM a transparent learning model. Another important feature of ssEAM is the capability of detecting atypical patters during either its training or performance phase. The detection of such patterns is accomplished via the employment of a match-based criterion that decides to which degree a particular pattern matches the characteristics of an already formed category in ssEAM. Additionally, via the utilization of hyper-ellipsoidal categories, ssEAM can learn complex decision boundaries, that arise frequently in gene expression classification problems. Finally, ssEAM can be easily implemented.

Since the datasets consist of only a small number of samples, it is better to use the jackknife approach, which is also known as leave one out cross validation (LOOCV), to examine the performance of the classifier [11]. For a dataset with N samples, the classifier is trained N times. Each time, a different single sample is left out as the test point and the other N-1 samples are used to train the classifier. The prediction performance of the classifier is estimated by considering the average accuracy of the N cross-validation experiments.

Generally, microarray data are easily overfitted, which requires a prudent experiment design process in order to assess the performance of classifiers fairly [12-13]. Here, we utilize the strategy that separates gene selection from the LOOCV operation in order to overcome the effect of selection bias, which is caused by including the test samples in the process of gene selection [13]. For each LOOCV iteration, informative genes are ranked and chosen according

Table I. CLASSIFICATION ACCURACY FOR THE SRBCT DATA SET. GIVEN ARE THE PERCENT OF CORRECT CLASSIFICATION FOR 83 TUMOR SAMPLES WITH LOOCV.

SRBCT	Features (Genes)						
	10	25	50	100	200	500	2308 (all)
EAM	97.6%	100%	100%	100%	100%	98.8%	87.9%
ssEAM (optimum ε)	98.8% (0.1, 0.2)	100% (0.1, 0.2, 0.3)	100% (0.1, 0.2, 0.3)	100% (0.1, 0.2, 0.3)	100% (0.1, 0.2)	98.8% (0.1)	87.9% (0.1)

Table II. CLASSIFICATION ACCURACY FOR THE NCI60 DATA SET. GIVEN ARE THE PERCENT OF CORRECT CLASSIFICATION FOR 58 TUMOR SAMPLES WITH LOOCV.

NCI60	Features (Genes)						
	10	25	50	100	200	500	1409 (all)
EAM	43.1%	70.7%	75.9%	77.6%	75.9%	72.4%	74%
ssEAM (optimum ε)	50% (0.2)	70.7% (0.1)	75.9% (0.1)	77.6% (0.1)	77.6% (0.1)	75.9% (0.1)	81% (0.1)

to the *N-1* samples with the Fisher discriminant criterion, described as

$$D(i) = \frac{|\mu_+(i) - \mu_-(i)|^2}{\sigma_+^2(i) + \sigma_-^2(i)} \qquad (2)$$

where $\mu_+(i)$ and $\mu_-(i)$ are the mean values of gene i for the samples in class +1 and class -1, and $\sigma_+^2(i)$ and $\sigma_-^2(i)$ are the variances of gene i for the samples in class +1 and -1. Therefore, the subsets of genes selected at each stage tend to be different. The Fisher score used here aims to maximize the between-class difference and minimize the within-class spread. Therefore, it gives the highest score to the gene that expresses itself most differently within two classes. Since our ultimate goal is to classify multiple types of cancer, we utilize a one-versus-all strategy to seek gene predictors.

III. Results

We test and analyze ssEAM performance in multiple cancer classification on the following two datasets. The first dataset is on the diagnostic research of small round blue-cell tumors (SRBCTs) of childhood and consists of 83 samples from four categories, known as Burkitt lymphomas (BL), the Ewing family of tumors (EWS), neuroblastoma (NB) and rhabdomyosarcoma (RMS) [5]. Gene expression levels of 2,308 genes are used in this analysis. The second dataset (NCI60) includes 1,416 gene expression profiles for 60 cell lines in a drug discovery screen by the National Cancer Institute [6]. These cell lines belong to 9 different classes: 8 breast (BR), 6 central nervous system (CNS), 7 colorectal (CO), 6 leukemia (LE), 9 lung (LC), 8 melanoma (ME), 6 ovarian (OV), 2 prostate (PR), and 8 renal (RE). Since the PR class only has two samples, they are excluded from further analysis.

Table I describes the best classification accuracy for the SRBCT dataset with the selection of different numbers of genes. ssEAM can achieve 100% accuracy when the number of selected gene predictors is in the range 25-200, which is consistent with the results obtained by other classifiers [5, 14]. The classification rate decreases to 87.9% at ε=0.1 when all genes are used. Likewise, the performance is deteriorated when only 10 or fewer genes are included in the subset. These results reflect the importance of gene selection in the context of tumor classification. Many genes are not related to the discrimination of certain cancer types of interest and including them in the dataset will bring noise into the classification system. On the other hand, important information will be wrongly discarded if inadequate genes are selected.

The results for the NCI60 dataset are summarized in Table II. In contrast with the results for the SRBCT datasets, the best performance (81%, better than other known results) is obtained when all genes are used at ε=0.1. In other words, the dimensionality reduction deteriorates the performance of classifiers instead of leading to an improvement as before. This is similar to the result reported by Berrar *et al.* [16], where no obvious improvement is observed for the reduced dataset. The reason may lie in the fact that some of the important genes cannot be effectively identified by the Fisher criterion. Currently, we also use a new evolutionary computation technique for feature selection, and can achieve better results [15]. We find that there is only a small fraction of overlaps between genes chosen by the two methods [15].

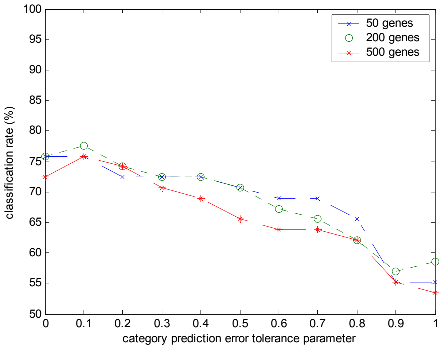

Fig. 3. The effect of the category prediction error tolerance parameter on classification rate (percent correct classification).

Fig. 3 shows the effect of the category prediction error tolerance parameter ε with respect to the classification accuracy. This strategy provides us an effective method to deal with over-fitting and increase the generalization of the classifier. Together with the results summarized in Table I and II, we can see that the performance of the classifier can usually be improved with an appropriate selection of the value of ε. The improvement is more evident when more genes are provided as inputs, in which a higher overlap among categories may exist.

IV. Conclusion

Cancer classification is critically important for cancer diagnosis and treatment. Microarray technologies provide a new and effective avenue to discriminating different kinds of cancer types, while simultaneously bringing many new challenges. Here, we utilized Semi-supervised Ellipsoid ARTMAP, combined with a simple dimensionality reduction strategy, to distinguish tumor tissues with more than two categories through analyzing gene expression profiling. Although the proposed method, together with other classifiers, has achieved qualitatively good results, there are still many problems to be solved, particularly, the curse of dimensionality, which is caused by the rapidly and persistently increasing capability of gene chip technologies, in contrast to the existing limitations in conditions like sample collections. This makes the published datasets consist of only a small set of samples for each tumor type, however, along with tens of thousands of gene expression measurements. Without any doubt, more samples are greatly helpful in effectively evaluating different kinds of classifiers and constructing cancer discrimination system. In the meantime, more advanced gene selection approaches are required in order to find informative genes that are relevant to the prediction and prognosis.

Acknowledgment

Partial support for this research from the National Science Foundation, and from the M.K. Finley Missouri endowment, is gratefully acknowledged.

References

[1] M. Eisen, and P. Brown, “DNA arrays for analysis of gene expression,” *Methods Enzymol*, vol. 303, pp. 179-205, 1999.

[2] R. Lipshutz, S. Fodor, T. Gingeras, and D. Lockhart, “High density synthetic oligonucleotide arrays,” *Nature Genetics*, vol. 21, pp. 20-24, 1999.

[3] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, “Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, pp. 531-537, 1999.

[4] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lander, and T. Golub, “Multiclass cancer diagnosis using tumor gene expression signatures,” *Proc. Natl. Acad. Sci. USA* 98, pp. 15149-15154, 2001.

[5] J. Khan, J. Wei, M. Ringnér, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, and P. Meltzer, “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks,” *Nature Medicine*, vol. 7, pp. 673-679, 2001.

[6] U. Scherf, D. Ross, M. Waltham, L. Smith, J. Lee, L. Tanabe, K. Kohn, W. Reinhold, T. Myers, D. Andrews, D. Scudiero, M. Eisen, E. Sausville, Y. Pommier, D. Botstein, P. Brown, and J. Weinstein, “A gene expression database for the molecular pharmacology of cancer,” *Nature Genetics*, vol. 24, pp. 236-44, 2000.

[7] G. Anagnostopoulos, M. Georgiopoulos, S. Verzi, and G. Heileman, “Reducing generalization error and category proliferation in Ellipsoid ARTMAP via tunable misclassification error tolerance: Boosted Ellipsoid ARTMAP,” *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN ’02)*, vol. 3, pp. 2650-2655, 2002.

[8] S. Grossberg, “Adaptive pattern recognition and universal encoding II: feedback, expectation, olfaction, and illusions,” *Biological Cybernetics*, vol. 23, pp. 187-202, 1976.

[9] G. Anagnostopoulos, and M. Georgiopoulos, “Ellipsoid ART and ARTMAP for incremental unsupervised and supervised learning,” *Proceedings of the IEEE-INNS-ENNS Intl. Joint Conf. on Neural Networks (IJCNN’01)*, vol. 2, pp. 1221-1226, 2001.

[10] G. Carpenter, S. Grossberg, N. Markuzon, J. Reynolds, and D. Rosen, “Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps,” *IEEE Transactions on Neural Networks*, vol. 3, pp. 698-713, 1992.

[11] R. Duda, P. Hart, and D. Stork, *Pattern classification*, 2nd Ed.. Wiley & Sons, New York, 2001.

[12] C. Ambroise, and G. McLachlan, “Selection bias in gene extraction on the basis of microarray gene-expression data,” *Proc. Natl. Acad. Sci. USA* 99, pp. 6562-6566, 2002.

[13] D. Nguyen, and D. Rocke, “Multi-class cancer classification via partial least squares with gene expression profiles,” *Bioinformatics*, vol. 18, pp. 1216-1226, 2002.

[14] Y. Lee, and C. Lee, “Classification of multiple cancer types by multicategory support vector machines using gene expression data,” *Bioinformatics*, vol. 19, pp. 1132-1139, 2003.

[15] R. Xu, G. Anagnostopoulos, and D. Wunsch, “Multi-class cancer classification using semi-supervised ellipsoid ARTMAP and particle swarm optimization with gene expression data,” in preparation.

[16] D. Berrar, C. Downes, and W. Dubitzky, “Multiclass cancer classification using gene expression profiling and probabilistic neural networks,” *Pacific Symposium on Biocomputing* 8, pp. 5-16, 2003.