Active Learning with Wavelets for Microarray Data[*]

D. Vogiatzis and N. Tsapatsoulis

Dept. of Computer Science, University of Cyprus,
Kallipoleos 75, Nicosia CY-1678, Cyprus
Tel:+357-22892749 Fax:+357-22892701
{dimitrv, nicolast@cs.ucy.ac.cy}

Abstract. In Supervised Learning it is assumed that is straightforward to obtained labeled data. However, in reality labeled data can be scarce or expensive to obtain. Active Learning (AL) is a way to deal with the above problem by asking for the labels of the most "informative" data points. We propose a novel AL method based on wavelet analysis, which pertains especially to the large number of dimensions (i.e. examined genes) of microarray experiments. DNA Microarray expression experiments permit the systematic study of the correlation of the expression of thousands of genes. We have applied our method on such data sets with encouraging results. In particular we studied data sets concerning: Small Round Blue Cell Tumours (4 types), Leukemia (2 types) and Lung Cancer (2 types).

keywords: active learning, wavelets, microarray data

1 Introduction to Active Learning

The idea that a large set of labeled data is available for training a classifier under a supervised regime is often wrong. Data labeling can be a time consuming and expensive process. For instance in the microarray experiments in biology, a large data set is produced which represents the expression of hundreds or thousands genes (i.e. production of RNA) under different experimental conditions (see [1] for an overview of microarray experiments), labeling exhaustively all the experimental samples can be very expensive or simply impossible. An alternative would be to start with a small set of labeled data, which may be easy to obtain, then to train a classifier with supervised learning. At this point a query mechanism pro-actively asks for the labels of some of the unlabeled data; whence the name active learning. The query implements a strategy to discover the labels of the most "informative" data points.

[*] Acknowledgment: The work presented in this paper has been undertaken in the framework of the OPTOPOIHSH project (PLHRO/0104/04 - Development of knowledge-based Visual Attention models for Perceptual Video Coding) funded by the Cyprus Research Promotion Foundation, Framework Programme for Research and Technological Development 2003-2005.

The concept of Active Learning is hardly new, an important contribution is in [2], where optimal data selection techniques for feedforward neural networks are discussed. In addition the authors show how the same techniques can be used for mixtures of Gaussians and locally weighted regression. An information based approach for active data selection is presented in [3]. In particular three different techniques for maximising the information gain are tested on an interpolation problem. In yet another approach, the geometry of the learning space is derived by computing the Voronoi tessellation, and the queries request the labels of data points at the borders of Voronoi regions [4]. The concept of active learning has also been realised in the context of Support Vector Machines for text classification in [5]. The method is based on selecting for labeling, data points that reduce the version space (the hyperplanes that separate the data) as much as possible. In sect. 2 we discuss about the use of wavelet analysis in bioinformatics, then in sect. 3 we present our proposed algorithm on wavelet based active learning. In sect. 4 we present the experimental setting and the results. In sect. 5 we elaborate on certain choices we made and especially, in the use of wavelets. Finally, in sect. 6 we present conclusions and future work.

2 Wavelets as data mining in Bioinformatics

Wavelets have been widely used in signal processing for more than 20 years. Moreover, their usefulness has also been proved in the domain of data mining [6] as well as in the biomedical domain [7]. The main advantage of the wavelets with respect to the Fourier transforms, is that they allow the localisation of a signal in both the time and frequency domains. From the point of view of mathematics, a function can be represented as an infinite series expansion in terms of a dilated and translated version of a basis function called the *mother wavelet*. For practical purposes, we can use the discrete wavelet transform, which removes some of the redundancy found in the continuous transform. In the experiments that we consider we have a small number vectors (microarray experiments), where the dimensions of each vector (genes involved) are an order of magnitude greater than the number of vectors (see [8] for overview of computational methods for microarray data). In particular a data set from a microarray experiment has the form of a two dimensional matrix $\mathbf{X}$. Let, $\underline{x}_i$ be a row vector of $\mathbf{X}$, and $\underline{x}_i = (x_{i,1}, x_{i,2} \dots x_{i,L})$. The index i refers to the microarray experiment or time step and its maximum value is relatively small, up to 100. On the other hand each dimension represents the expression level (i.e. the RNA that has been produced) of a specific gene in a specific experiment. A characteristic property of microarray which differentiates them from most of the other data is that the number of samples is small when compared to the number of dimensions. Because of this property each sample can be considered as a time series.

3 Wavelet based Active Learning

It is presumed that we have a pool of labeled and unlabeled multidimensional data, and that the number of dimensions is far greater than the number of data. The purpose is to reduce the error of a classifier (built with a labeled training set), by selectively asking for the labels of the unlabeled data.

We propose an algorithm which implements a query strategy based on wavelet analysis, and it can be informally stated as: find two data items, one from the testing set of the classifier and the other from the pool of unlabeled data, such that their distance is minimal and the item from the testing set is from the worst performing class (in terms of the Root Mean Square Error). Perform that computation on multiple levels of wavelet analysis, which possibly returns multiple candidates from the unlabeled pool. Then apply a voting scheme and choose the best unlabeled data item from the pool. Next, we present formally the algorithm:

1. Train the classifier with the labeled data.
2. Perform 1D, discrete wavelet transform on each labeled and unlabeled data item, from scale S_1 up to $S_{log_2(L)}$, where L is the number of dimensions. Here greater index in scale denotes lower resolution of wavelet analysis and S_1 denotes the original signal.
3. *Active Learning step:* Form a query to ask for the label of an unlabeled datum: Let class k is the class with the worst classification error and let $\mathbf{X_k}$ be the set of data classified, by the current classifier, to class k. Let also $\underline{x}_{k,i} \in \mathcal{R}^L$ be the i-th signal in $\mathbf{X_k}$. By $\underline{x}_{k,i}^{S_r}$ we denote signal $\underline{x}_{k,i}$ at scale S_r, $r \in \{\mathit{0}, .., \mathit{log_2}(L)\}$. If $\mathbf{U}$ is the set of unlabeled data and $\underline{u}_j \in \mathcal{R}^L$, is the j-th signal in $\mathbf{U}$, represented at scale S_r by $\underline{u}_j^{S_r}$, then the following criterion is used to select the next candidate datum $\underline{u}_\xi$ at scale S_r:

$$\xi^{S_r} = \arg\min_j \{\min_i \|\underline{u}_j^{S_r} - \underline{x}_{k,i}^{S_r}\|_2\} \tag{1}$$

   where $\underline{u}_j^{S_r}$ denotes that the comparison is made at scale S_r. It is possible (though not necessary) that $\underline{u}_\xi^{S_r} \neq \underline{u}_\xi^{S_l}$ $r, l \in \{\mathit{0}, .., \mathit{log_2}(L)\}$, $r \neq l$.
4. The previous step is performed for all S_r, which produces a series of

$$\underline{u}_\xi^{S_1} \dots \underline{u}_\xi^{S_{log_2(L)}} \tag{2}$$

5. Assign weights to all $\underline{u}_\xi^{S_r}$, where each weight is reversely proportional to the eucledian distance of $\underline{u}_\xi^{S_r}$ from its closest $\underline{x}_{k,i}^{S_r}$
6. All $\underline{u}_\xi^{S_1} \dots \underline{u}_\xi^{S_{log_2(L)}}$ vote according to a weighted majority voting scheme.
7. The label of winner from the previous voting $\underline{u}_\xi^{S_r}$ is requested, and the datum is entered into the training data set, and the classifier is retrained with the updated training set.
8. The algorithm terminates, when the user decides that the overall classifier's performance is good enough, or when there can be no more labels.

The rationale of the algorithm, is to query for labels to the data that are closest to the worst performing data in terms of the classification error. Asking at multiple scales, implies that the similarity that it is sought must be present at multiple resolutions; and thus more impervious to noise.

4 Experiments

We have used datasets from 3 labeled microarray experiments. The first data set was obtained from "The Microarray Project cDNA Library" (http://research.nhgri.nih.gov/microarray/Supplement/). The second and third data sets were obtained from the Gene Expression Datasets collection (http://sdmc.lit.org.sg/GEDatasets).

The first data set is about Small Round Blue-Cell tumours (SRBCT), investigated with cDNA microarrays containing 2308 genes, over a series of 63 experiments. The 63 samples included tumour biopsy material and cell lines from 4 different types: 23 Ewing's sarcoma (EWS), 20 rhabdomyo sarcoma (RMS), 12 neuroblastoma (NB) and 8 Burkitt's lymphoma (BL). There are also available 20 samples (6 EWS, 3 BL, 6 NB and 5 RMS) for testing [9].

The provenance of the second data set stems is also from oligonucleotide microarrays, with a view of distinguishing between acute lymphoblastics leukemia (ALL) and acute meyeloid leukemia (AML). The training data set consisted of 38 bone morrow samples (27 ALL, 11 AML) from 7130 human genes. The test data set consisted of 34 samples (20 ALL, 14 AML) [10].

The third data set also stems from a microarray experiment and consists of lung malignant pleural mesothylioma (MPM) and adenocarcinoma (ADCA) samples [11]. The training set consists of 32 samples (16 MPM and 16 ADCA) each class) from 12534 human genes. The test set consists of 149 samples (15 MPM and 134 ADCA).

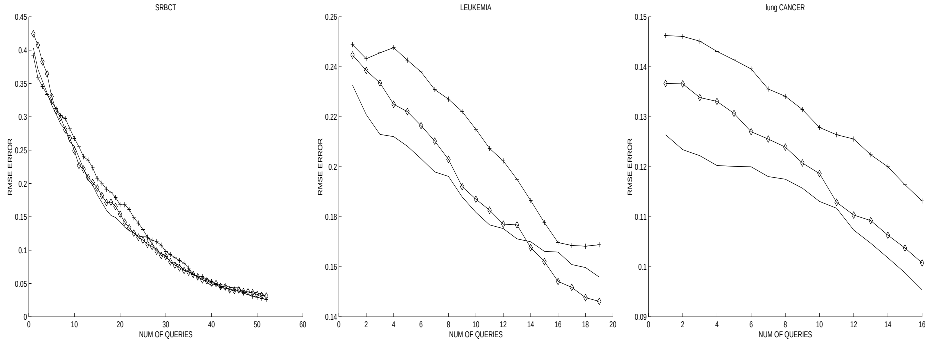

Fig. 1. ◇: (diamond line) random choice of datum, __: (solid line) wavelet weighted majority voting, +-+ (cross line) minimum distance at the largest scale

He have split the data sets into training, pool and testing subsets. The first is used for training the classifier. Under the active learning regime a query asks for the labels of specific data from the pool. The testing subset is used for independent control (see Table 1). The best datum (according to the query asked) receives a label and it is subsequently integrated into the training data set. The results reported in Fig. 1 depict the testing set error and are averages over 100 experiments. In each experiment the whole data set is shuffled while the number of training, pool and test sets remains the same.

As a classifier, we have used a Support Vector Machine (with a polynomial kernel of degree 3), primarily for practical reasons; our smallest data set has 2308 dimensions and training a multilayered perceptron would take a prohibitively long time. Training on all data sets always resulted in learning 100% of the training sample.

Table 1. Data Sets characteristics and Performance

Description					Cumulative Query Error		
Data set	Training	Pool	Test	# dimensions	Random	Distance	Wavelet
SRBC	13	52	20	2308	7.4655	7.7540	7.2860
Leukemia	19	19	34	7130	3.6385	4.0050	3.5556
Lung	16	16	117	12534	1.9301	2.1144	1.8165

The experimental setting aims to compare the reduction of the classification error on the testing set by comparing three different query types. Each query type progressively asks for the labels of all data in the pool. Next, we present the three query types (the Wavalet Distance Query is what we proposed):

Random Query For comparison only: The next datum $\underline{u}_\xi$ to be included in the training data set is selected from the pool at random.

Distance Query The selected datum (from the pool) is the one which minimises: $\|\underline{u}_\xi - \underline{x}_{k,i}\|$, without applying any wavelet transformation. k denotes the category where the classifier has the highest Root Mean Square Error (RMSE), i is the index of the datum $\underline{x}$, and $\underline{x}$ is selected from the testing subset.

Wavelet Distance Query The proposed method: Apply the criterion of distance query at all scales of wavelet analysis. The winner from each scale votes according to a weight that is reversely proportional to the minimum distance. We have applied a 1D discreet wavelet (Daubechies family order 2) transform up to scale $log_2(L)$, where L is the number of dimensions and $\underline{x}$ is selected from the testing subset.

The results are summarised in Fig. 1 and in Table 1 where the cumulative query error is reported over all pool data for all three data sets (lower values are better). For all data sets the wavelet based active learning method outperformed the two other methods.

All experiments were carried out on the Matlab 6.5 platform, with the OSU SVM classifier and the Wavelet toolboxes.

5 Discussion

The signals we consider are made of the expression levels of genes. The task we are addressing is that of supervised learning, by active (on the part of classifier) selection of the training samples. The signal is not assumed to be stationary, thus the frequencies extend over limited regions of the signal. The wavelet transform offers the best trade off between localisation and extraction of frequencies. Let us assume that the wavelet transform was not applied, and the best datum was selected from the pool for labeling, based on the proximity to another datum of known label. Proximity, can be defined in terms of euclidean/mahalanobis etc. distance. This would impose a certain structure on data belonging in the same category (i.e. categories are spheres/ellipsoids) which of course is not necessarily the case. Naturally, data of the same category must have something in common, but this common property has to been "mined". With this work, we advocate that the intrinsic properties of each datum are hidden in its frequencies and their location. This analysis has to proceed at different scales, because higher frequencies tend to be shorter in duration than lower frequencies, subsequently lower frequencies can be resolved in frequency.

To enforce the claim that wavelet analysis is necessary to detect the most similar object; i.e. it is better to look at multiple scales than at a single scale, it is interesting to observe the data of Fig. 1, where the selection based on minimum distance at largest scale is even worse than selecting at random.

Principal Component Analysis (PCA) is widely used in statistics and machine learning. In many cases the data dimensions are highly correlated with each other; PCA can transform the dataset in such a way that the new dimensions are uncorrelated, then the new dimensions with the lowest variance could be discarded. The assumption of PCA is that the intrinsic data dimensionality is lower that the original dimensionality. In the problem domain we have considered each dimension is the expression level of a gene, and whereas some of the dimensions are expected to be related there is no guarantee that they are linearly related. Concomitant to that is that if the data distribution is not gaussian like, PCA being a linear transformation will not be of much use. Therefore, a better choice in the vein of PCA is to implement and experiment with non-linear extensions of PCA.

6 Conclusions and Future work

We have designed a method for active learning with a weighted wavelet based voting scheme and we evaluated the method on three datasets from microarray experiments with encouraging results —the wavelet distance query outperformed all other methods. It is interesting to observe that the simple distance query performs on average worse than random choice.

We intent to apply our proposed method on a larger number of microarray data sets to test experimentally its validity. In particular the Gene Expression Datasets collection contains a large number of microarray data sets (see http://sdmc.lit.org.sg/GEDatasets).

An important issue is the choice of the unlabeled data item to be included in the training set, in our case we have devised a weighted voting scheme, however there are also other good choices such as considering only the scale with the lowest entropy which have to be investigated.

We also need to investigate the role of the wavelet. In particular we have use the Daubechies order 2 (db2)—an orthogonal wavelet. What would be the result of a biorthogonal wavelet? It remains to test it experimentally. However, in the experiments we conducted the Haar wavelet underperformed when compared to db2. Moreover, the characteristics of the microarray experiments, might lead us to design a new wavelet to fit the problem of active learning. Another important issue is that the order of the components in the data vectors is not important, they represent expression levels of genes, therefore a rearrangement of the components might lead to improved results.

Finally, we have not taken advantage of the characteristics of the Support Vector Machine in particular we have not considered the separating hyperplanes that are produced. This could suggest a way of improving the active learning query.

References

1. Bergeron, B.: Chapter 6. In: Bioinformatics Computing. Prentice Hall (2004) 222–231
2. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. In Tesauro, G., Touretzky, D., Leen, T., eds.: Advances in Neural Information Processing Systems. Volume 7., The MIT Press (1995) 705–712
3. MacKay, D.: Information-based objective functions for active data selection. Neural Computation **4** (1992) 590–604
4. Hasenjäger, M., Ritter, H.: Active learning with local models. Neural Processing Letters **7** (1998) 107–117
5. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. In Langley, P., ed.: Proceedings of ICML-00, 17th International Conference on Machine Learning, Stanford, US, Morgan Kaufmann Publishers, San Francisco, US (2000) 999–1006
6. Li, T., Li, Q., Zhu, S., Ogihara, M.: A survey on wavelet applications in data mining. SIGKDD Explorations **4** (2003) 49–68
7. Liò, P.: Wavelets in bioinformatics and computational biology: state of art and perspectives. Bioinformatics **19** (2003) 2–9
8. Quackenbush, J.: Computational Analysis of Microarray Data. Nature Reviews **2** (2001) 418–428
9. Khan, J., Wei, J., Ringer, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C., Meltzer, P.: Classification and diagnostic

prediction of cancers using gene expression profiling and artificial neural network. Nature Medicine **7** (2001) 673–679

10. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science (1999)
11. Gordon, G., Jensen, R., Hsiao, L., Gullans, S., Blumenstock, J., Ramaswamy, S., Richard, W., Sugarbaker, D., Bueno, R.: Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Research (2002) 4963–4967