High Throughput Analysis of Breast Cancer Specimens on the Grid

Lin Yang[1,2], Wenjin Chen[2], Peter Meer[1], Gratian Salaru[2], Michael D. Feldman[3], and David J. Foran[2]

[1] Dept. of Electrical and Computer Eng., Rutgers Univ., Piscataway, NJ, 08544, USA
[2] Center of Biomedical Imaging and Informatics, The Cancer Institute of New Jersey, UMDNJ-Robert Wood Johnson Medical School, Piscataway, NJ, 08854, USA
[3] Dept. of Surgical Pathology, Univ. of Pennsylvania, Philadelphia, PA, 19104, USA

Abstract. Breast cancer accounts for about 30% of all cancers and 15% of all cancer deaths in women in the United States. Advances in computer assisted diagnosis (CAD) holds promise for early detecting and staging disease progression. In this paper we introduce a Grid-enabled CAD to perform automatic analysis of imaged histopathology breast tissue specimens. More than 100,000 digitized samples (1200×1200 pixels) have already been processed on the Grid. We have analyzed results for 3744 breast tissue samples, which were originated from four different institutions using diaminobenzidine (DAB) and hematoxylin staining. Both linear and nonlinear dimension reduction techniques are compared, and the best one (ISOMAP) was applied to reduce the dimensionality of the features. The experimental results show that the Gentle Boosting using an eight node CART decision tree as the weak learner provides the best result for classification. The algorithm has an accuracy of 86.02% using only 20% of the specimens as the training set.

1 Introduction

Breast cancer is a malignant neoplasm that can affect both women and men. It is the leading cancer in both white and African American women, with more than 178,480 new cases for an estimated to be diagnosed in 2007 where 2030 cases are men, and will be responsible for estimated 40,460 deaths [1]. It is the second most common cause of cancer death in white, black, Asian/Pacific Islander and American Indian/Alaska Native women [1,2]. The incidence of breast cancer among women has increased gradually from one in 20 in 1960 to one in eight today. At this time there are slightly over 2 million breast cancer survivors in the United States. Women living in North America have the highest rate of breast cancer in the world [1].

In spite of the increase in the incidence of the disease, the death rates of breast cancer continue to decline. This decrease is believed to be the result of earlier detection through screening and analysis as well as improved treatment [1]. Some of the common methods screening of breast cancer include examination by a physician, self-performed routine examinations and routine mammograms. When suspicious lesions are detected by these methods, a fine needle aspirate or

N. Ayache, S. Ourselin, A. Maeder (Eds.): MICCAI 2007, Part I, LNCS 4791, pp. 617–625, 2007.

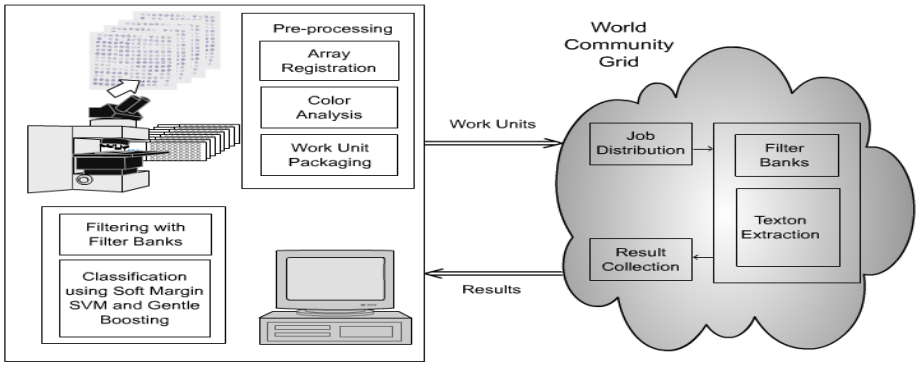

Fig. 1. The work-flow and logical units of the Grid-enabled tissue microarray

biopsy can be performed and the obtained tissue is examined [3]. The extracted tissue is mounted on glass slides and examined by a surgical pathologist to make decision of benign or cancer. Diaminobenzidine (DAB) and hematoxylin are standard staining methods used for breast cancer.

There has been increasing interest in investigating computer assisted diagnosis (CAD) system to help breast cancer diagnosis, e.g. [4]. However, to our knowledge, most of these studies were relatively limited in scale because the computation is often a bottleneck. In this paper, we report a Grid-enabled framework for analyzing imaged breast tissue specimens. We discriminate between benign and cancer breast tissues using the results obtained from the Grid. Without using the Grid, the filtering and generation of universal texture library would require 210 days of computation for 3744 samples because of the large image size and computational complexity of the process.

2 Grid Enabled Tissue Microarray and Features

Tissue Microarray (TMA) technology, e.g. [5], provides a platform to perform comparative cancer studies involving evaluation of protein expression. Several commercial products have been developed to automate the process of digitizing TMA specimens, e.g. T2 scanner, Aperio Technologies and MedMicro, Trestle Corporation. With the recent advance of the Grid technology [6,7], by now we can address the computational complexity of large-scale collaborative applications by leveraging aggregated bandwidth, computational power and secondary storage resources distributed across multiple sites. IBM has offered their World Community Grid to help research projects which require high levels of computation. Our Grid-enabled system was launched on the IBM World Community Grid in July, 2006. The work-flow and logical units are shown in Figure 1.

Textons are defined as repetitive local features that humans perceive as being discriminative between textures. We used the 49×49 LM filter bank [8] composed of 48 filters: eight LOG filter responses with $\sigma = 1, \sqrt{2}, 2, 2\sqrt{2}, 3, 3\sqrt{2}, 6, 6\sqrt{2}$,

four Gaussian filtering responses with $\sigma = 1, \sqrt{2}, 2, 2\sqrt{2}$ and the bar and edge filtering response within six different directions, $\theta = 0, \pi/6, \pi/3, \pi/2, 2\pi/3, 5\pi/6$, $\sigma = 1, \sqrt{2}$, 2. All the σ-s are in pixel size. While textons are used to describe the appearance of breast cancer images, they are not specific for our applications.

The image filtering responses obtained from the Grid were collected together, and clustered using K-means, $K = 4000$ to capture a large enough code book. The texton library was constructed from the cluster centers. The appearance of each breast tissue image was modeled by a compact quantized description - texton histogram, where each pixel is assigned to its closest texton using

$$h(i) = \sum_{j \in I} count(T(j) = i) \tag{1}$$

and I denotes breast tissue image, i is the i-th element of the texton dictionary, $T(j)$ returns the texton assigned to pixel j. In this way, each breast tissue image is mapped to a point in the high dimension space R^d, where d is equal to the number of textons. Figure 2 shows the texton histograms of two imaged breast tissue specimens, one benign (top-left) and one cancer (bottom-left).

3 Dimension Reduction and Classification

Each breast cancer image is represented by a vector in the $d = 4000$ dimension space in which we have to consider the "curse of dimensionality". Traditionally, linear dimension reduction methods like multidimensional scaling (MDS) and principle component analysis (PCA) [9] have been used, which assume that the data can be represented by a lower dimensional linear subspace. In other cases, the data can be modeled by a low-dimensional nonlinear manifold, and nonlinear methods such as locally linear embedding (LLE) [10], isometric feature mapping (ISOMAP) [11] and local tangent space alignment (LTSA) [12] are more appropriate.

3.1 Dimension Reduction

Given a set of feature vector $Z = \{z_1, ...z_i, ..., z_n\}$ where $z_i \in R^d$. There exists a mapping T which can represent z_i in the low dimension as

$$z_i = T(x_i) + u_i \quad i = 1, 2, ..., n \tag{2}$$

where $u_i \in R^d$ is the sampling noise and $x_i \in R^{d'}$ denotes the representation of z_i in low-dimensional space.

PCA finds the mapping T which best represents the variance of data z_i in the original high-dimensional space. The low-dimensional space $R^{d'}$ is spanned by the d' largest eigenvectors of the covariance matrix of z_i. MDS finds the mapping T which preserves the pairwise distances between z_i-s in $R^{d'}$. If the data have certain geometric structure which can be modeled as a low-dimensional linear manifold, they are not expected to perform well.

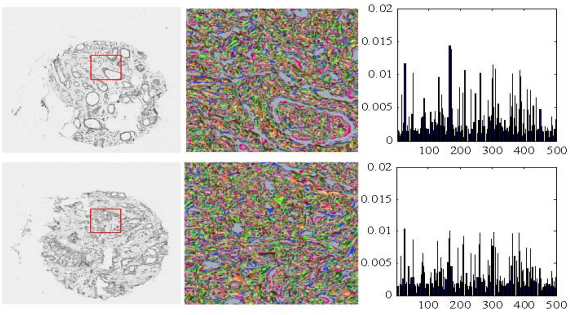

Fig. 2. Two breast tissue images on the left, benign (on the top) cancer (on the bottom). In order to keep the figure readable, only the texton maps of red rectangle regions were shown in the middle with different colors representing different textons. In practice the texton histograms of full images, shown on the right, were used for classification.

Although most of the geometric distances can not be used directly on the manifold, linear dimension reduction methods can be applied locally and Euclidean distances are valid on the local tangent space. The LLE preserves the geometry of the data points by representing points using their local neighbors. The ISOMAP approach applies a linear MDS on the local patch, but aim to preserve the geometric distance globally using the shortest path in the graph. The LTSA method maps each data point from the original space into the tangent space and align it to give a global coordinate.

If the data are approximately sampled from a low-dimensional manifold, the nonlinear methods generally perform better [10]. Otherwise the linear methods may be preferred because of their simplicity. In order to compare the performance of these methods, we used cross-validation (CV)

$$CV(\gamma) = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - f^{-k(i)}(x_i, \gamma)\right| \tag{3}$$

where x_i is the feature vector in $R^{d'}$, $y_i = \{+1, -1\}$ represents the cancer and benign breast tissue labels. The $f^{-k}(x_i, \gamma)$ denotes the classification results using the γ-th dimension reduction method with the k-th part removed from the training data, and the number of partitioning $k = 5$ in our experiments. We tested the five different algorithms from 3 to 4000 dimensions and determined that ISOMAP is better (Figure 3). Good performance using ISOMAP was also reported in [13] for tissue characterization. Based upon these comparative results, ISOMAP with 500 dimensions are used for all subsequent operations.

3.2 Classification

In [14], KNN and $C4.5$ decision tree were integrated into a Bayesian classifier which provided good results for characterizing breast tissues. In [15], a cascade boosting classifier is used to detect prostate cancers. In our case, each breast

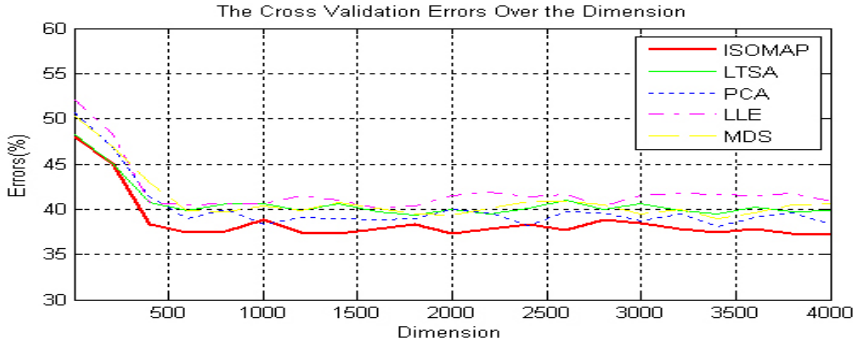

Fig. 3. The five different dimension reduction algorithms shown as five-fold cross-validation errors for different reduced dimensions

Input: Given n features $x_i \in R^{d'}$ and their corresponding labels $y_i = \{-1, 1\}$.
Training:
- Pick up 100 random pairs. Compute
  $\tau = \frac{1}{100} \sum_{i,j=1}^{100} \chi^2(x_i, x_j)$ where $\chi^2(x_i, x_j) = \frac{1}{2} \sum_{l=1}^{500} \frac{\left(x_i(l) - x_j(l)\right)^2}{x_i(l) + x_j(l)}$.
- Build the Mercer kernel $\kappa(x_i, x_j) = \exp(-\frac{1}{\tau}\chi^2(x_i, x_j))$.
- Select the penalty parameter C and train the soft margin $SVM(\kappa, C)$. Record the model parameters.

Testing:
- Output the classification: $sign\left[SVM(x)\right] = sign\left[\sum_i y_i \alpha_i \kappa(x, x_i) + b\right]$ where α_i and b are the learned weights and learned threshold.

Alg. 1. Soft Margin SVM using the Mercer kernel based on χ^2 distance

tissue image is represented by a feature vector in the reduced subspace $x_i \in R^{d'}$, where $d' = 500$. The maximum margin classifiers such as Support Vector Machine (SVM) [16] and Boosting [17] are more appropriate, especially when the number of training vectors are comparable with their dimensions.

Soft Margin Support Vector Machine. Because the training data are not linearly separable, we choose the Soft Margin SVM which allows training vectors to be on the wrong side of the support vector classifier with certain penalty. We use a nonlinear Mercer kernel [18] based on χ^2 distance. The detailed description is given in Algorithm 1.

The key parameters which affect the accuracy of soft margin SVM are the penalty and the kernel. The penalty parameter C can be selected according to cross validation (CV) errors. For the kernel selection, we tested linear, polynomial, Gaussian and the proposed Mercer kernel based on χ^2 distance, where the last one outperformed all the others.

Gentle Boosting on the Decision Tree. Boosting is one of the most important recent developments in machine learning. It works by sequentially applying

Input: Given n features $x_i \in R^{d'}$ and their corresponding labels $y_i = \{-1, 1\}$.
Training:
- Initialize the weights $w_i = 1/n, i = 1, ..., n$. Set $b(x) = 0$ and the number of nodes $M = 8$ in the CART decision tree.
- For $j = 1...J$
  - Each training sample is assigned its weight w_i. The weighted tree growing algorithm is applied to build the CART decision tree $T_j(x, M)$.
  - Update using $b(x) = b(x) + T_j(x, M)$.
  - Update the weights $w_i = w_i e^{-y_i T_j(x,M)}$ and renormalize w_i.
  - Save the j-th CART decision tree $T_j(x, M)$.

Testing:
- Output the classification: $sign\left[b(x)\right] = sign\left[\sum_{j=1}^{J} T_j(x, M)\right]$.

Alg. 2. Gentle Boosting using an eight nodes Classification And Regression Tree (CART) decision tree as weak learner

a classification algorithm on a reweighted version of the training data and the final label is decided by a weighted voting. Instead of using the most well-know Adaboost [17] with a simple linear classifier, as the weak classifier, we propose to apply Gentle Boosting [19] using an eight node CART decision tree as the weak learner, which experimentally provided higher accuracy than Adaboost. The detailed algorithm is provided in Algorithm 2.

The number of nodes of the CART decision tree can be selected using cross validation (CV). In our experiments, using eight nodes CART decision tree as weak learner provided the best results. The number of iterations J was chosen as 40 to achieve satisfactory accuracy and avoid overfitting as well.

4 Experiments

The tissue microarrays used in our experiments are prepared by different institutes: the Cancer Institute of New Jersey, Yale University, University of Pennsylvania and Imgenex Corporation, San Diego, CA. To date over 300 immunostained microscopic specimens, each containing hundreds of tissue image, were digitized at 40× volume scan using the Trestle MedMicro, a whole slide scanner system. The output images typically contain a few billions of pixels and are stored as a compressed tiled TIFF file sized at about two gigabytes. The registration protocol proposed by [20] was applied to automatically identify the rows and columns of the tissue arrays. Staining maps of the two dyes, diaminobenzidine (DAB) and hematoxylin, were generated from specimens and each of the two staining maps as well as the luminance of the original color image were submitted to the IBM World Community Grid for batch processing.

We have analyzed 3744 breast cancer tissues (674 hematoxylin and 3070 DAB staining) from the 100,000 images processed on the Grid. Without the Grid, it would require about 210 days of computation to generate the texton library even with an efficient C++ implementation on a PC with P3 1.5GHz processor and

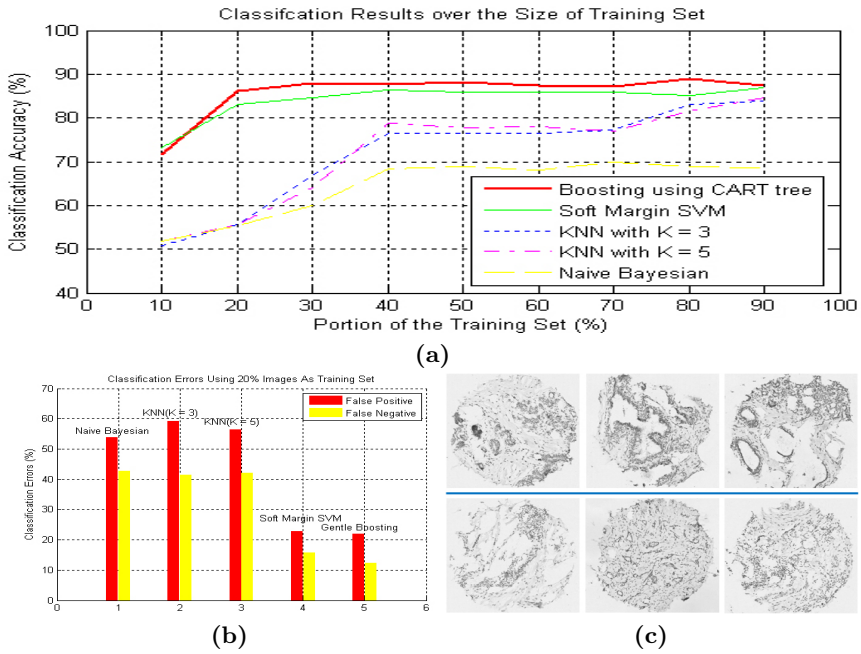

Fig. 4. The classification results. (a) Accuracy as the function of the size of training set using Gentle Boosting, Soft Margin SVM, KNN with $K = 3$ or 5 and naive Bayesian. (b) The false positive and false negative errors using 20% images as training set. (c) Some misclassified samples. The upper row is false positive and lower row is false negative.

1G RAM. However, we can build this universal texton library in less than 40 minutes in the largely distributed computing system [6].

The labels of all breast tissues from the hospitals and institutions are independently confirmed by certificated surgical pathologists. The dataset used in these experiments consisted of 611 benign and 3133 cancer samples. Each of the five algorithms was applied 10 times, using different parts of the training images drawn by random sampling. Figure 4 shows the average classification results. Because there were more positive samples than the negative samples, we obtained higher false positive errors but lower false negative errors (Figure 4b) than the average error (Figure 4a). It is clear that Gentle Boosting and Soft Margin SVM performed better, especially when the training set is small. The overall best one is the Gentle Boosting using an eight node CART decision tree as the weak learner, in which case the classification accuracy is 86.16% using only 20% of the dataset for training.

5 Conclusion

We have presented a Grid-enabled framework using texture features to perform high throughput analysis of imaged breast cancer specimens. A Gentle Boosting

using an eight node CART decision tree as the weak learner provided the best results. In our experiments Atypical Duct Hyperplasia (ADH) gave highest false positive rates. In the future, we plan to subclassify the stages of breast cancer progression using the universal texton library, which were generated from the clustering results returned from the IBM World Community Grid.

Acknowledgements

This research was funded, in part, by grants from the NIH through contract 5R01LM007455-03 from the National Library of Medicine and 5R01EB003587-02 from the National Institute of Biomedical Imaging and Bioengineering. We would like to thank IBM's World Community Grid support team and our collaborators at The Cancer Institute of New Jersey and University of Pennsylvania.

References

1. American Cancer Society: Cancer Facts and Figures 2007. 2007 edn. American Cancer Society (2007)
2. U. S. Cancer Statistics Working Group: United states cancer statistics: 2003 incidence and mortality (preliminary data). National Vital Statistics 53(5) (2004)
3. Rosai, J.: Rosai and Ackerman's Surgical Pathology. 9th edn. Mosby (2004)
4. Suri, J.S., Rangayyan, R.M.: Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer. 1st edn. SPIE (2006)
5. Hoos, A., Cordon-Cardo, C.: Tissue microarray profiling of cancer specimens and cell lines: Opportunities and limitations. Mod. Pathol. 81(10), 1331–1338 (2001)
6. Berman, F., Fox, G., Hey, A.J.G.: Grid Computing: Making the Global Infrastructure a Reality, 1st edn. Wiley, Chichester (2003)
7. Egan, G.F., Liu, W., Soh, W.S., Hang, D.: Australian neuroinformatics research - grid computing and e-research. ICNC 1, 1057–1064 (2005)
8. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. IJCV 43(1), 29–44 (2001)
9. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley, Chichester (2000)
10. Rowels, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
11. Tenebaum, J., de Silva, V., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
12. Zha, H., Zhang, Z.: Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM J. on Sci. Comp. 26(1), 313–338 (2004)
13. Lekadir, K., Elson, D.S., Requejo-Isidro, J., Dunsby, C., McGinty, J., Galletly, N., Stamp, G., French, P.M., Yang, G.Z.: Tissue characterization using dimensionality reduction and fluorescence imaging. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4191, pp. 586–593. Springer, Heidelberg (2006)
14. Oliver, A., Freixenet, J., Marti, R., Zwiggelaar, R.: A comparison of breast tissue classification techniques. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4191, pp. 872–879. Springer, Heidelberg (2006)

15. Doyle, S., Madabhushi, A., Feldman, M., Tomaszeweski, J.: A boosting cascade for automated detection of prostate cancer from digitized histology. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4191, pp. 504–511. Springer, Heidelberg (2006)
16. Cortes, C., Vapnik, V.: Support vector networks. Mach. Learn. 20, 1–25 (1995)
17. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of online learning and an application to boosting. J. Comp. and Sys. Sci. 55(1), 119–139 (1997)
18. Shawe-Taylor, J., Cristianini, N.: Kernel Mathods for Pattern Analysis, 1st edn. Cambridge University Press, Cambridge (2004)
19. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. Machine Learning, 148–156 (1996)
20. Chen, W., Reiss, M., Foran, D.J.: Unsupervised tissue microarray analysis for cancer research and diagnosis. IEEE Trans. Info. Tech. on Bio. 8(2), 89–96 (2004)