Class Discovery in Colorectal Cancer Gene Microarray Data

Will Cukierski
Pattern Recognition, Spring 2007
May 9, 2007

Abstract: Gene microarrays have been studied as a means to find relevant genes within a large set of candidates. This selection is typically accomplished through clustering algorithms that utilize a variety of statistical similarity measures. As opposed to gene selection, this study explores a colorectal cancer gene microarray to find clinically-meaningful groups of patients (as judged from pathology reports). The existence of a pre-cancerous class is investigated. Several methods of dimensionality reduction, classification, and unsupervised clustering are performed and compared within the framework of gene microarrays.

1 Background

Research on cancer microarrays straddles two fields of inquiry. The clinical side concentrates on finding genes and genomic features with biological implications. The algorithmic side uses microarrays as high-dimensional data on which to compare machine-learning algorithms. Hundreds of informatics studies have examined specific genes which express differently in tumor vs. normal tissue [1], yet researchers still lack a comprehensive, genomic description of the tumor process. Cancer informatics has made significant progress towards the two-class classification problem, as well as the refinement of significant gene features. Classification algorithms can work with or without domain-specific knowledge and learn the most relevant features through training. Modern classification algorithms acting on gene microarrays can typically delineate (> 80% accuracy) the two-class problem for various forms of cancer [2]. There is also an interest, however, to mine gene microarrays to find new, non-obvious associations within the data. Several studies have applied unsupervised clustering methods to microarrays [3, 4, 5, 6, 7], yet much remains to be studied within this paradigm.

Approximately 2/3 of the published literature on cancer microarrays focuses on class discovery [8]. This study employs several unsupervised clustering algorithms on a much-expanded colorectal cancer dataset from Alon et al. [7]. The data set consists of 22,283 mRNA features from an Affymetrix oligonucleotide array. Collection details are given in [7, 9]. 360 tissue samples of six tissue types were included (Table 1). Abbreviations of the classes are as follows: CC = colon cancer, PP = primary polyp, NM = normal mucosa, LIM = liver metastasis, LUM = lung metastasis, NLI = normal liver, NLU = normal lung, T1CC = T1 graded colon cancer ...

Tissue Type   # Samples   Mean Correlation
Colorectal Cancer 174 0.92
Primary Polyp 50 0.94
Liver Metastasis 47 0.91
Lung Metastasis 20 0.93
Normal Mucosa 49 0.92
Normal Liver 13 0.94
Normal Lung 7 0.96

Table 1: Correlation coefficients are Pearson coefficients, $r = \frac{\text{cov}(x, y)}{\text{std}(x) \cdot \text{std}(y)}$.

The study will seek evidence of multiple classes (possibly corresponding to precancerous states) and examine the clustering results in lower-dimensional representations. The results are compared to pathologist reports on the sampled tissue specimens.

2 Preliminaries

Before clustering the microarray data, it is helpful to first examine the basic structure of the data. As one might biologically expect, the gene profiles are highly correlated, especially within the same tissue type (see Tab. 1 and Fig. 1). High correlation poses both a practical and theoretical challenge to analysis by statistical techniques. Practically, one desires high intra-class correlation with low inter-class correlation. In the colorectal dataset, the intra-class correlation (91-96%) and inter-class (88%) are quite close. In crude terms, any clustering/classification algorithm must exploit the narrow range of the few percentage points which separate these two. The theoretical challenge of high correlation is the selection of features which give rise to the narrow differences between classes. Fig. 1 shows a subtle feature of gene expression data. If one removes genes with the highest variance, the histogram shifts towards uncorrelation. This decrease suggests the difference between samples is the result of many subtle expression changes in genes with little variance in expression...
intensity. As shown by Alon et al., gene expression levels may also be highly correlated with other genes [7]. It is statistically tempting, but biologically naive, to assume the uncorrelated genes (those that may give rise to class separation) are implicated in the pathological process.

3 Dimensionality Reduction

The high dimensionality of gene microarrays introduces the well-known curse of dimensionality. While it is possible to lower dimensionality through manual selection of genes, this work takes a domain-inspecific approach: all the collected features are used in an unsupervised fashion. This approach reduces the possibility of selection bias, whereby features are chosen to produce a pre-conceived result/classification, but entails one drawback. As stated in Notterman et al., samples of distinct tissue types will have different tissue composition [9]. A biopsy of benign colon tissue, for example, is expected to contain more muscle tissue than a biopsy of cancer. There exists the danger that a clustering or classification algorithm exploits such "contaminations" to separate classes. The effects of these biases can be reduced, in an unsupervised fashion, by removal of the features with the most variance. By using all features indiscriminately, we assume the intra-class variance between samples of the same tissue type is the sole result of the disease process.

Nonlinear methods of dimensionality reduction have recently been shown to be better suited to microarray data [10]. To confirm this result, Principle component analysis (PCA), a linear technique which seeks to preserve the variance in the high-dimensional data, was compared with the isometric feature mapping algorithm (ISOMAP) of Tenenbaum et al. [11]. The superiority of nonlinear methods is qualitatively verified in Fig. 2. ISOMAP clearly retains a global ordering which is lost in the PCA embedding. The space of microarray data is extremely sparse (the number of samples is much less than the dimensionality). Since the mRNA features are highly correlated, the data inhabits a vast, unorthogonal space in which Euclidean distance has limited meaning. Instead, as accomplished by ISOMAP, one seeks a geodesic metric that accounts for nonlinearity. The failure of PCA to preserve nonlinear, global structure suggests that use of linear distance metrics, whether in clustering, classification, or dimensionality reduction, will fail to aid in class discovery.

The primary cell of origin is also visible in the embedded dataset (Fig. 3). Biopsies taken from liver and lung metastases are embedded with samples of the primary colorectal

---

1 In addition to tissue composition, secondary biological changes associated with hyperplasia, but not specific to cancer, may bias the classification.

2 The mean Euclidean distance between data points is $5.9 \times 10^4$ units in the original 22,283 dimensional space. By comparison, the vector $(1, 1, \ldots, 1)$ has length 149.2 units.
Figure 2: 2D embedding of the colon cancer, polyp, and normal groups. ISOMAP clearly embeds the data, showing the transition from normal to polyp to cancer. PCA fails to separate the classes. Labels are the ground truth from the pathology reports.

(a) Metastases are embedded separate from the host tissue.

(b) Liver metastases are embedded with the cancer group, but clearly on the side of the normal liver samples.

Figure 3: Cell of origin is visible in ISOMAP embedding.
tumor. The embedding of the liver metastases, in particular, shows a genetic profile which is colorectal in origin, but partially reflects the host environment (the samples are embedded between the cancer points and the normal liver points). It is unclear whether this effect is produced by contamination of the biopsy with small amounts of normal tissue, or genetic adaptation of the metastasis to the host environment.

### 4 Classification

Supervised classification was performed as a precursor to unsupervised clustering. If a properly trained classifier cannot predict clinical stage, for example, the data may not reflect this particular partition. A polynomial Support Vector Machine (SVM) [12] classifier was trained using 1/3 holdout cross validation (repeated 100 times) on several partitions of the dataset (Fig. 2). The classifier readily separated the normal vs. cancer two-class problem and the metastatic lesion from the host tissue. The classifier performed poorly, however, in distinguishing the clinical stage and Crohn’s reaction partitions. It is also reported here that the same classifier was trained using leave-one-out cross validation (results not shown). With this training, the classifier was able to separate all classes, including clinical stage, with negligible error. These optimistic error estimates are assumed to be the result of overtraining in the high-dimensional space and not indicative of successful classification.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Dim.</th>
<th>Err(%)</th>
<th>Var(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>NM</td>
<td>-</td>
<td>-</td>
<td>22283</td>
<td>1.37</td>
<td>0.01</td>
</tr>
<tr>
<td>CC</td>
<td>NM</td>
<td>PP</td>
<td>-</td>
<td>22283</td>
<td>9.74</td>
<td>0.07</td>
</tr>
<tr>
<td>CC</td>
<td>NM</td>
<td>PP</td>
<td>LIM</td>
<td>22283</td>
<td>10.96</td>
<td>0.07</td>
</tr>
<tr>
<td>NLI</td>
<td>LIM</td>
<td>-</td>
<td>-</td>
<td>22283</td>
<td>2.32</td>
<td>0.14</td>
</tr>
<tr>
<td>NLU</td>
<td>LUM</td>
<td>-</td>
<td>-</td>
<td>22283</td>
<td>0.13</td>
<td>0.02</td>
</tr>
<tr>
<td>T1CC</td>
<td>T2CC</td>
<td>T3CC</td>
<td>T4CC</td>
<td>22283</td>
<td>45.98</td>
<td>0.30</td>
</tr>
<tr>
<td>T2CC</td>
<td>T4CC</td>
<td>-</td>
<td>-</td>
<td>22283</td>
<td>35.06</td>
<td>1.18</td>
</tr>
<tr>
<td>CC</td>
<td>LIM</td>
<td>LUM</td>
<td>-</td>
<td>22283</td>
<td>7.80</td>
<td>0.12</td>
</tr>
<tr>
<td>CRY</td>
<td>CRN</td>
<td>-</td>
<td>-</td>
<td>22283</td>
<td>31.53</td>
<td>0.26</td>
</tr>
<tr>
<td>CC</td>
<td>NM</td>
<td>-</td>
<td>-</td>
<td>10</td>
<td>3.53</td>
<td>0.03</td>
</tr>
<tr>
<td>CC</td>
<td>NM</td>
<td>PP</td>
<td>-</td>
<td>10</td>
<td>14.94</td>
<td>0.12</td>
</tr>
<tr>
<td>CC</td>
<td>NM</td>
<td>PP</td>
<td>LIM</td>
<td>10</td>
<td>21.50</td>
<td>0.10</td>
</tr>
<tr>
<td>NLI</td>
<td>LIM</td>
<td>-</td>
<td>-</td>
<td>10</td>
<td>1.05</td>
<td>0.11</td>
</tr>
<tr>
<td>NLU</td>
<td>LUM</td>
<td>-</td>
<td>-</td>
<td>10</td>
<td>4.13</td>
<td>0.35</td>
</tr>
<tr>
<td>T1CC</td>
<td>T2CC</td>
<td>T3CC</td>
<td>T4CC</td>
<td>10</td>
<td>53.66</td>
<td>0.38</td>
</tr>
<tr>
<td>T2CC</td>
<td>T4CC</td>
<td>-</td>
<td>-</td>
<td>10</td>
<td>45.50</td>
<td>1.37</td>
</tr>
<tr>
<td>CC</td>
<td>LIM</td>
<td>LUM</td>
<td>-</td>
<td>10</td>
<td>18.48</td>
<td>0.12</td>
</tr>
<tr>
<td>CRY</td>
<td>CRN</td>
<td>-</td>
<td>-</td>
<td>10</td>
<td>32.42</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Table 3: k-means clusters, in general, do not correspond to classes of tissue labels, clinical stage, or Crohn’s test. Only in cases of the uncorrelated, normal liver samples does k-means partition according to tissue type.

Hierarchical clustering is the most widely-used, unsupervised method for gene microarrays [8]. A hierarchical classification method was implemented, again using correlation as the distance metric. A hierarchical tree was constructed using the weighted pair group method using arithmetic averaging (WPGMA). The method initializes each sample as its own cluster, and merges the clusters with the highest correlation. The resulting clusters can be visualized in a dendrogram, as shown in Fig. 5.

Lastly, a novel method of clustering by Horn et al. [6], called Quantum Clustering (QC), was also implemented. This physics-inspired method creates a probabilistic wave function and potential function, which constitute a solution to the Schrödinger equation. Singular value decomposition is first performed, then each data point is assigned to a Gaussian of width $\sigma$ by a Parzen-window approach,

$$\psi(\vec{x}) = \sum_i e^{-\frac{(x-x_i)^2}{2\sigma^2}}$$

A "potential" is then created according to,

$$V(\vec{x}) = -\min \left( \frac{\sigma^2 \nabla^2 \psi}{2\psi} + \frac{\sigma^2 \nabla^2 \psi}{2\psi} \right)$$

where the "min" term makes $V$ positive. The minima of $V$

3Defined as the ratio of the correlation between samples within a cluster to the correlation with samples in other clusters. For sample $x_i$ in cluster $A$,

$$\frac{\sum_{j \in A} \text{corr}(x_i, x_j)}{\sum_{j \notin A} \text{corr}(x_i, x_j)}$$
(a) k-means with $k = 2$ partitions normal and metastatic liver samples according to class type, empirical error = 0.033.

(b) k-means with $k = 3$ fails to partition colon cancer samples into meaningful classes, empirical error = 0.355.

Figure 4

(c) The dendrogram of the clusters above shows partitioning of the data into five tree branches.

Figure 5
are then the cluster centers. One free parameter is varied and indirectly determines the number of clusters. The performance of the QC algorithm is illustrated in Fig. 6.

\[ -\frac{\sigma^2}{2} \nabla^2 + V \psi = E \psi \]

contains this information, and secondly, that the clustering algorithm discovers such a class. As shown in the 2D, ISOMAP embedding of Fig. 2, there appears to be a trajectory on the nonlinear manifold that captures the transition from normal to cancerous. The same trajectory is not seen, however, under the premise that the clinical staging reflects the time course of the cancer. Notterman et al. also observed no relationship between clinical stage and the gene profile [9].

The next step in this work involves the comparison of cluster results against the pathology reports, as well as statistical validation of the clusters. It is hoped that the methods return clusters which correlate with relevant clinical features. In particular, further attempts will be made to correlate clinical stage with the genetic profile.

### 6 Discussion

The discovery of a pre-malignant class is thus predicated on two assumptions: firstly, that the microarray actually

\[ \text{contains this information, and secondly, that the clustering} \]

\[ \text{algorithm discovers such a class. As shown in the 2D,} \]

\[ \text{ISOMAP embedding of Fig. 2, there appears to be a} \]

\[ \text{trajectory on the nonlinear manifold that captures the transition} \]

\[ \text{from normal to cancerous. The same trajectory is not seen,} \]

\[ \text{however, under the premise that the clinical staging reflects} \]

\[ \text{the time course of the cancer. Notterman et al. also observed} \]

\[ \text{no relationship between clinical stage and the gene profile} \]

\[ [9]. \]

\[ \text{The next step in this work involves the comparison of cluster} \]

\[ \text{results against the pathology reports, as well as statistical} \]

\[ \text{validation of the clusters. It is hoped that the methods return} \]

\[ \text{clusters which correlate with relevant clinical features. In} \]

\[ \text{particular, further attempts will be made to correlate clinical} \]

\[ \text{stage with the genetic profile.} \]

### References


