A Novel Approach for Classifying Gene Expression Data using Topic Modeling

Soon Jye Kho\textsuperscript{*}
Knoesis Center, Wright State University
3640 Colonel Glenn Hwy
Dayton, Ohio 45431
soonjye@knoesis.org

Michael L. Raymer
Knoesis Center, Wright State University
3640 Colonel Glenn Hwy
Dayton, Ohio 45431
michael.raymer@wright.edu

Hima Bindu Yalamanchili\textsuperscript{*}
Knoesis Center, Wright State University
3640 Colonel Glenn Hwy
Dayton, Ohio 45431
himay.87@gmail.com

Amit P. Sheth
Knoesis Center, Wright State University
3640 Colonel Glenn Hwy
Dayton, Ohio 45431
amit@knoesis.org

1 INTRODUCTION

Traditional diagnosis for cancer is based on clinical and morphological data, but these methods have been reported to have limitations in their diagnostic ability.\cite{1, 12}. To overcome the limitations, cancer detection based on genomic data has been proposed\cite{17, 18}. In recent years, with the wide employment of microarray and next-generation sequencing methods, increase in data volume poses both promise and challenges to researchers in identifying patterns and analyzing the data\cite{15}. Although there has been a lot of research in the identification of differentially expressed genes associated with various types of cancer tissues, typically small sample size and high dimensionality of expression data continues to pose challenges to researchers. Understanding and interpretation of results remains a key challenge in analyzing gene expression data.

Topic modeling, a machine learning approach has shown promise in the fields of text mining and image retrieval, where it has been successfully implemented to extract information from high dimensional data\cite{6, 14, 24}. Latent Dirichlet Allocation (LDA) is one of the most popular topic modeling approaches in text mining among others like Latent Semantic Indexing (LSI) and Probabilistic Latent Semantic Analysis (PLSA)\cite{9, 11}. Since its emergence, researchers have implemented this approach in biomedical text mining as well\cite{4, 25}.

Given the successful implementation of topic models in discovering the useful structure of the documents, researchers are beginning to implement topic modeling approaches to analyze data other than document collections\cite{8, 27}. Recently, there have been efforts to use topic modeling techniques in the field of bioinformatics to perform unsupervised analysis and obtain insights into high-dimensional -omics data\cite{13, 16, 22, 23}. To gain better understanding into cancer classification and gene identification from such data, in this paper, we focus on the class prediction of breast cancer and lung cancer based on gene expression data using topic modeling.

We approached this study in three phases (Figure 1). In the first phase, we employed LDA as a clustering application. LDA has been successfully applied as a clustering algorithm for gene expression data in the past\cite{28}. Here, we confirmed that LDA-derived document clusters based on our rank-based data transformation technique show clear separation between cancer and healthy tissue...
samples. In the second phase, we investigated the application of LDA as a classification algorithm. This involved optimizing several algorithmic parameters including the number of topics and algorithm passes with respect to classification of gene expression data. We used 10-fold cross validation to estimate classification effectiveness. We compared our results with alternative machine learning techniques including Support Vector Machine (SVM), Naive Bayes and Random Forest classification, as well as Principal Component Analysis (PCA) and Hierarchical Clustering for unsupervised learning. In the third phase, we additionally investigated the effectiveness of LDA in identifying genes differentially regulated between cancer and normal tissues. We explored the degree to which specific topics identified by the LDA algorithm correspond to gene regulatory pathways which vary between tumor and normal tissues.

The remaining of this paper is organized as follows. In Section 2, we review the related work of implementing LDA in gene expression data. In Section 3, we describe data collection and preprocessing while in Section 4 we describe our proposed method, evaluation techniques, results from two data sets and interpretation of identified differentially expressed genes.

2 RELATED WORK

One of the challenges in using LDA for genetic studies is the nature of the data. The input of topic modeling is typically a simplified bag-of-words representation of a corpus. However, gene expression data is numerical in nature, and thus needs to be transformed into text to provide appropriate input for the LDA algorithm. The most commonly used transformation method is scaling the matrix and interpreting the discrete values as gene/word occurrences. The higher the expression value, the higher the frequency of the gene in the bag of words [3, 19]. One limitation using this approach is genes that have zero expression value will not be present in the training corpus. Consequently, this transformation approach is suitable for clustering and feature reduction, but not for classification because the test set might have genes that are unseen by the trained model.

More recently, Zhao et al. [28] used a different transformation method for generating a corpus. For each gene, expression values were normalized to 0 (lower than median) or 1 (higher than median). The samples were then transformed into a bag-of-words, containing only genes with normalized value 1. This transformation method does not capture the deviation between samples. That is, two differentially expressed genes are treated the same regardless of the degree of deviation. Besides, this method is not feasible for imbalanced dataset as the cutoff point between two different classes could not be represented by median value. Rogers et al. [22] used Latent Process Decomposition (LPD), a derivative of LDA, to capture the continuous nature of gene expression data. No significant difference in terms of accuracy has been reported using LPD compared to classical approaches like pLSA [3].

To alleviate these limitations and broaden the range of applications using LDA, we propose a novel transformation strategy to generate bag of words. We adapted the gene ranking method from Wang et al. [26], which is described in detail in Section 3.2. The rationale behind this ranking is that relative ordering of gene expression is generally stable in a particular type of normal human tissue but is widely disturbed in a diseased tissue. Ranking method is able to capture the degree of deviation of gene expression across different samples and still enable gene interpretability of topics using topic profiles.

3 METHODS

3.1 Data Collection and Filtering

We obtained the mRNA-Seq data of 229 breast invasive carcinoma samples and 98 lung squamous cell carcinoma samples from The Cancer Genome Atlas (TCGA) data portal 1. We artificially balanced the data using only the tumor samples with matched normals. Out of 229 breast cancer samples, 117 represent primary solid tumor and 112 represent solid normal tissue. Of the 98 lung cancer samples, half of the data (49) represent primary solid tumor and the other half (49) represent solid tissue normal. Both datasets have 60483 variables with an abundance of zero values – roughly 45% of the datasets. Genes that have zero expression value in more than 10% of samples were removed. This filtering step might remove some differentially expressed genes but such stringent threshold was used to prevent any bias (multiple bins having genes with exclusively zero expression value) that might arise in the preprocessing step. After filtering, a total of 23424 genes remained in the breast cancer data and 23996 genes remained in the lung cancer data.

3.2 Preprocessing

Here, we used a ranking approach to transform gene expression values into a bag-of-words. Within each sample, genes were sorted in ascending order based on their expression values. Expression values form a distribution that is divided into 20 bins. The performance of the algorithm is robust against the bin size (5, 10, 20, 30, 40 bins) and this is expected since the number of bins represents the relative ranking of expression level instead of absolute ranking (data not shown).

We chose 20 bins as a representative of the performed experiments. Actual expression values were then replaced with a quantized value (expression rank) based on its position. Value 1 indicates the group with lowest expression level and value 20 indicates the group with highest expression level. The combination of each gene with its corresponding expression rank (separated by a hyphen), generated a bag of words for each sample. For example, if gene BRCA1 was found in quantized bin 3 for a particular sample, the bag-of-words associated with that sample would include the word BRAC1-3 to represent the ranked and quantized expression of this gene. It is important to note that as a result of quantization, comparison between gene and word may not be absolute since one gene might represent many words across the samples. To extract gene-specific information from the LDA-derived model, it is thus necessary to develop a technique to reverse this one-to-many mapping and trace back genes from the corresponding words without deviating much from the expression rank (Section 4.2). After the transformation, the breast cancer data corpus had 229 documents and 23424 words within each document. The lung cancer data corpus had 98 documents and each document contained 23996 words.

1https://portal.gdc.cancer.gov/
4 RESULTS AND ANALYSIS

For each dataset, LDA was implemented using an open-source Python library, Gensim [21]. The number of passes and number of topics were set to 10 and 3 respectively. These parameters were determined experimentally to provide the optimum performance (data not shown). The LDA algorithm produces two matrices: the topic-sample matrix, which expresses the computed probability of topics in a particular sample (document), and the topic-word matrix, containing the probability distribution over all available words for each topic. The LDA-derived matrices were utilized to perform analysis in 3 phases: clustering, classification and gene interpretation.

4.1 Clustering

Average topic distributions of samples from each class (normal and cancer) are depicted in Figure 2. This topic-sample matrix can be used for feature extraction/projection in a manner analogous to principal component analysis (PCA). Each column vector of the topic-sample matrix represents a single sample. There is one value in this vector for each topic, expressing the degree of association between the topic and the sample. Complete-linkage and euclidean-distance-based Hierarchical Clustering was applied on the topic-sample matrix to cluster the samples with similar probability distributions. The heatmap in Figure 2 shows that for both breast cancer and lung cancer data, samples were clearly grouped together into two classes, cancer and normal. Colors from red to green in the heatmap represent topic probabilities of the samples, ranging from 0 to 1.

The clustering result from LDA was compared with other conventional clustering and projection methods including hierarchical clustering and PCA. Hierarchical clustering was directly applied to the raw expression values for breast cancer (23424 genes) and lung cancer (23996 genes). PCA was first used to transform the original data into components that retain 90% of the variance. Hierarchical clustering was then applied on the transformed data. For both datasets, cluster purity and number of misclassified samples were calculated using Equation 1 to evaluate the clustering performance.

\[
Purity = \frac{1}{N} \sum_{i=0}^{k} \max_{j} |c_i \cap t_j|
\]  

(1)

Where \( N \) = number of data points, \( k \) = number of clusters, \( c_i \) = a cluster in K, \( t_j \) = classification that has the maximum count for cluster \( c_i \).

Table 1 shows the clustering results of the three methods where the number of clusters (K) is fixed at 2, 3, and 4 clusters. Overall, LDA clusters align more closely to known cancer/healthy tissue labels than those obtained by PCA and Hierarchical Clustering. One advantage of LDA-based clustering for gene expression data may be its mutually non-exclusive assumption. Most non-fuzzy clustering approaches have a mutually exclusive/independence assumption that a sample/gene is restricted to only one cluster. This assumption might not be logical for gene expression data largely when a sample/gene share characteristics with more than one cluster (that is one gene can be involved in different pathways). Thus using LDA for clustering of gene data would reflect the complex interplay between genes and pathways and improve quality of the results.
The lower the MSE, the more similar the topic.

In this section, we propose a novel approach to classification using LDA-based topic modeling, and explore the effectiveness of the proposed method in distinguishing gene expression patterns in breast and lung cancer tissues from those in healthy tissue. In contrast to the work by [3] where LDA is solely used as a feature extraction method and requires running LDA on both training and testing data, the approach proposed here applies the LDA algorithm only to the training data.

First, training of the LDA algorithm proceeds in the same manner as described previously for LDA-based clustering. For testing, LDA-derived topic-sample and topic-word matrices from the trained model are employed. To revert/collapse the genes from their corresponding words, expression rank and probabilities of all words representing a gene are taken and normalized over the entire probability distribution (Equation 2):

\[
exp_g = \frac{\sum_{j=1}^{n_g} \text{rank}_j^g \times \text{prob}_j^g}{\sum_{j=1}^{n_g} \text{prob}_j^g} \quad (2)
\]

where \(exp_g\) represents the collapsed expression rank of a gene \(g\), \(j\) is the total number of words that representing gene \(g\). \(\text{rank}_j^g \in (\text{rank}_1^g, \text{rank}_2^g, ..., \text{rank}_{n_g}^g)\) is a vector of rank extracted from words and \(\text{prob}_j^g \in (\text{prob}_1^g, \text{prob}_2^g, ..., \text{prob}_{n_g}^g)\) is the vector of corresponding probabilities for the words.

Combining the topic-word probabilities in this manner results in a new topic-gene matrix, in which each topic is associated with a specific rank value for each gene. The resulting topic profile facilitates subsequent classification.

To classify a test sample based on the trained LDA, we need to determine topic probability distribution of the test sample. First, the same pre-processing step described for clustering is applied to the test data, resulting in a bag-of-words based on ranked gene expression for each test sample. Then similarity (MSE) between test data, resulting in a bag-of-words based on ranked gene expression for each test sample. Then similarity (MSE) between test data and each topic profile was calculated using Equation 3. The lower the MSE, the more similar the topic.

\[
MSE = \sum_{g=1}^{q} (exp_g - bin_g)^2 \quad (3)
\]

Where \(exp_g\) represents the collapsed expression rank of gene \(g\) in a topic profile and \(bin_g\) represents expression rank of gene \(g\) in test sample. Gaussian normalization was performed to reduce the range of the MSE, thus avoiding overflow errors in computation, without affecting the variation between topics and the softmax function (Equation 4) is applied to ensure that the probability distribution properly sums to 1.0 [5]. The probability distribution determined from similarity is used for prediction.

\[
\text{SoftmaxOutput} = \frac{e^{sMSE}}{\sum_{k=1}^{K} e^{sMSE}} \quad (4)
\]

The performance of our classification approach was compared with popular supervised methods including SVM, Naive Bayes and Random Forest classifiers using 10-fold cross validation. Traditional metrics for comparison like accuracy, precision, recall and F-measure were applied. As shown in Table 2, LDA achieves a competitive performance comparable to other algorithms.

### 4.3 Gene Interpretation

We utilized topic profiles from the two data sets to perform pathway analysis and disease annotation of differentially expressed genes within each topic. A list of differentially expressed genes (DEGs) is extracted by computing the difference in expression rank of each topic with a baseline. For the experiments shown here, topic 3 was chosen as the baseline in both breast cancer and lung cancer datasets, since it shows the highest probability in normal samples (Figure 2). DEGs were then identified from all other topics (1 and 2 in this case). A threshold rank-difference of 5 was used to extract the significant genes and also limit the number of genes for further analysis. Thus, the extracted genes have at least 5 ranking changes between normal and tumor that could potentially represent significant dysregulation in our analysis.

We examined the pathway enrichment of the differentially expressed genes using DAVID Bioinformatics Resources 6.8 [10] to explore the KEGG pathway database. Figures 3 and 4 show the number of genes and statistically enriched pathways in the topics for breast cancer and lung cancer, respectively.

The relevance of the identified differentially expressed genes was validated using the DAVID bioinformatics suite [10] and the

---

**Table 1: Clustering results of LDA-derived topic probabilities, PCA reduced features, and Hierarchical Clustering.**

<table>
<thead>
<tr>
<th>Method</th>
<th>K</th>
<th>Breast cancer</th>
<th>Lung cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No. of Misclassified Samples</td>
<td>Cluster Purity</td>
</tr>
<tr>
<td>Latent Dirichlet Allocation (LDA)</td>
<td>2</td>
<td>11</td>
<td>0.952</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11</td>
<td>0.952</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>11</td>
<td>0.952</td>
</tr>
<tr>
<td>PCA</td>
<td>2</td>
<td>112</td>
<td>0.512</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>112</td>
<td>0.512</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>109</td>
<td>0.524</td>
</tr>
<tr>
<td>Hierarchical Clustering</td>
<td>2</td>
<td>112</td>
<td>0.511</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>112</td>
<td>0.511</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>112</td>
<td>0.511</td>
</tr>
</tbody>
</table>

Where \(exp_g\) represents the collapsed expression rank of gene \(g\) in a topic profile and \(bin_g\) represents expression rank of gene \(g\) in test sample. Gaussian normalization was performed to reduce the range of the MSE, thus avoiding overflow errors in computation, without affecting the variation between topics and the softmax function (Equation 4) is applied to ensure that the probability distribution properly sums to 1.0 [5]. The probability distribution determined from similarity is used for prediction.

\[
\text{SoftmaxOutput} = \frac{e^{sMSE}}{\sum_{k=1}^{K} e^{sMSE}} \quad (4)
\]

The performance of our classification approach was compared with popular supervised methods including SVM, Naive Bayes and Random Forest classifiers using 10-fold cross validation. Traditional metrics for comparison like accuracy, precision, recall and F-measure were applied. As shown in Table 2, LDA achieves a competitive performance comparable to other algorithms.

---

2https://david.ncifcrf.gov/home.jsp
NIH genetic association database (GAD) [2]. The DEGs were extracted by comparing the expression profile for normal and tumor samples. For each gene, its expression rank ($exp_d$) in topic profile matrix (~24000×3) was multiplied with respective topic probability distribution of normal group (3×1) and tumor group (3×1). This would generate two expression profiles: one for normal samples (~24000×1) and one for tumor samples (~24000×1). The difference in expression rank between these two profiles was computed. Top 250 DEGs was extracted both for breast cancer and lung cancer and annotated using DAVID Bioinformatics Resources 6.8 in terms of adjusted $p$-value, the higher the significance of the pathway.

Table 3 shows the top 5 highly related diseases which are ranked ascendingly according to their adjusted $p$-value.

As expected, for the gene list extracted from breast cancer data corpus, ‘breast cancer’ is the disease with highest enrichment (lowest adjusted $p$-value) and for lung cancer data corpus, ‘smoking cessation’ and ‘chronic obstructive pulmonary disease’ are highly enriched.

5 DISCUSSION

Many approaches to diagnostic classification based on mRNA expression focus primarily on differential expression. The LDA-based approach described here differs in that the focus is primarily on co-expression. Just as textual LDA attempts to group co-occurring
words into topics in order to explain the topic composition of a document, gene expression LDA can be used to identify co-regulated groups of genes that together explain the overall patterns of gene expression in healthy and disease states. In an unsupervised mode, this technique has shown a somewhat surprising ability to produce gene-collections (topics) that differ significantly between cancer and healthy tissues. Augmenting this technique with class labels results in an even more capable diagnostic classifier.

Here, our approach is only tested on gene expression data from next generation sequencing. Nevertheless, we expect it to work equally with similar data from different platforms or for different tasks such as disease subtype classification [7], survival analysis and treatment prognosis prediction [20]. We have shown that by using only gene expression data, the model was able to cluster the data into topics that are biologically coherent and meaningful. An interesting next step would be to include additional lifestyle and clinical metadata in the bag-of-words representing each sample. For example, by tagging clinical study data with labels such as smoker/non-smoker, male/female, adult/child, etc., associations between subject groups corresponding to these labels, disease state, and gene expression might further be identified. Understanding associated pathways and genomic etiology in these tasks is important in stratifying patients according to their risk and provide better diagnosis.

6 CONCLUSION

Overall our approach provides a novel direction for applying the LDA algorithm to identify and group differentially expressed genes between healthy and cancer tissues of various types. A novel technique for transformation of gene expression levels to words is presented and shown to be effective. Comparative evaluation of this approach with state-of-the-art pattern classification methods confirms the effectiveness of the proposed methodology. Differential gene expression patterns associated with lung and breast cancer were identified and validated as relevant using pathway analysis and the NIH’s genetic association database. Next steps include testing the effectiveness of this LDA-based classification approach on genetic and epigenetic variation patterns and also testing the effectiveness of LDA as a supervised algorithm for the more complex problem of distinguishing among cancer subtypes.

ACKNOWLEDGMENTS

This work is partially funded by the Biomedical Sciences Ph.D. program at Wright State University.

REFERENCES