Diagnosis of Viral Infection Using Gene Expression Profiles

Harleen Billing
Northwestern University
harleen@u.northwestern.edu

Clifton McFate
Northwestern University
mcfateclifton79@gmail.com

Sahil Shah
Northwestern University
sahil@u.northwestern.edu

1. INTRODUCTION & MOTIVATION

According to the Centers for Disease Control and Prevention, respiratory syncytial virus (RSV) results in an average 177,000 hospitalizations and 14,000 deaths per year among adults older than 65 and influenza resulted in 64 pediatric deaths during the 2012-2013 flu season. [5],[6] A reliable method for early detection for respiratory infections like these could result in better health outcomes.

The gene expression levels of a cell correspond to the abundance of mRNA produced during transcription of each gene. These levels change upon viral infection. Thus, difference in gene expression profiles may be a good early indicator of illness.

However, using gene expression profiles for classification is very difficult because there are a large number of genes measured for any cell (10^4) and thus a huge range of possible gene profiles. While researchers have data on how these genes react to illness, they do not know which changes are meaningful. Thus the feature space is not well defined. There is also quite a bit of variability across individuals, which means any learning method needs to be robust to cross-subject variance. We propose and evaluate a k-nearest neighbor classifier for this task across multiple k-values and distance measures.

2. DATA

Our training set consisted of gene expression profiles from peripheral blood mononuclear cells (PBMCs) collected from patients with acute influenza infection, respiratory syncytial virus (RSV) infection, and neither influenza nor RSV infection to serve as a control group. This data was collected in [2]. The accession number of our dataset is GSE34205

Each example consists of a classification, infected or uninfected, along with a vector of gene expression levels. Each vector consists of 54,675 gene expression levels, which are real valued numbers. Samples were collected from 79 ill patients. A control group of uninfected cells were collected from 22 patients. In this data set, each patient is a data point. Thus we have a total of 101 instances to base our model on.

3. SYSTEM OVERVIEW

Our system uses principal component analysis to reduce the dimensionality of the gene profiles. For a new unlabeled cell, the system compares its profile to the database of classified profiles using L1 norm, L2 norm, or cosine similarity. The most common label among the top k scoring exemplars is the output label. This is either infected or uninfected.

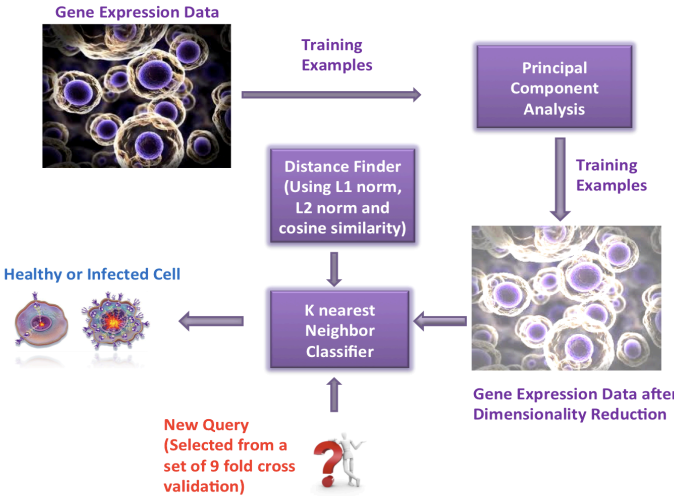

Figure 1: System Work-Flow

3.1 Feature Selection

Since there are usually tens of thousands of genes probed per sample in gene expression data, the first step in our system is to reduce the dimensionality of the data set. We do this by applying principal component analysis to our dataset and the selecting the top 25 components. This explains 83% of the variance in our data. From there, we select as our relevant genes those genes whose coefficient is greater than 2 standard deviations from the mean. This results in 20,279 genes, a 63% reduction in dimensionality.

4. EXPERIMENTS

The goal of the experiment is to determine which distance measures and k-values best classify infected cells. Using the above system:

We varied both the k-values and distance measures for the classifier and evaluated each combination using 9-fold validation.

- We have 101 examples. We performed 9-fold validation. 8 folds had 11 samples and 1 had 13.
- We tried k values: 1,3,5,7.
- We tried distance measures: L1 Norm, L2 Norm, cosine similarity
- For each test-fold, results were output as two cell arrays, one of predicted output and one of actual output. Each test fold also produced an accuracy score.

We then repeated each of these trials using the same randomly selected set of PCA components to evaluate the effect of our PCA approach. Running without PCA turned out to be too resource intensive to be practical.

Initial data analysis revealed that uninfected and infected cells were unevenly distributed throughout the data. This resulted in some folds containing the majority of uninfected cells, greatly limiting the number remaining in the training set. To resolve this, we randomized the order of the cells before creating the folds. The relative distributions of infected and uninfected cells across folds are shown below.

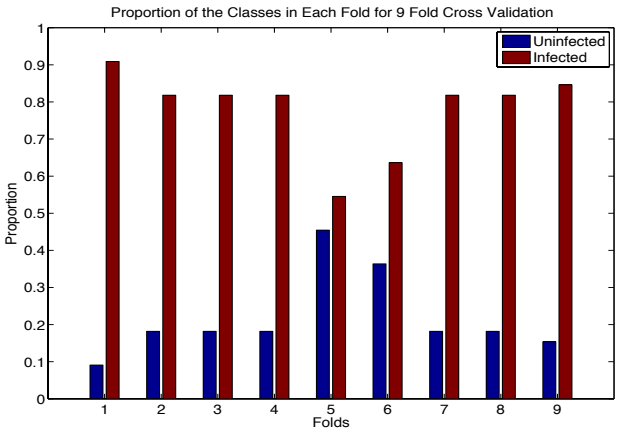

Figure 2: Proportion of Classes Across Folds

5. RESULTS

Our top performing method was a k-value of 1 and L1 norm at 92.075% accuracy. These results are summarized in table 1 below.

K value	L1 Norm	L2 Norm	Cosine
1	0.92075	0.9021	0.72028
3	0.892	0.892	0.71018
5	0.90909	0.90909	0.70008
7	0.88889	0.90909	0.71018

Table 1: Average Accuracy

For each result we then computed the average F1 measure which is a weighted ratio between the rates of precision and recall. An F score of 1 is perfect. This data is summarized in *table 2*.

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

K Value	L1 Norm	L2 Norm	Cosine
1	0.94452	0.93458	0.8301
3	0.92718	0.92718	0.82596
5	0.93974	0.93974	0.81953
7	0.92933	0.94415	0.82512

Table 2: Average F score

The F1 measure for our top scoring method was 0.9445. This was an increase of 15.25% in F1 from the lowest performing set (cosine similarity, k = 5). Against a baseline that always chose infected (Fscore = 0.8779) our method provided a 7.5% increase.

Furthermore, we evaluated our PCA strategy against a baseline that used random components. The top performing measure again was k = 1 with L1 norm. With an average F score of 0.944, there was no significant difference between the two methods. Interestingly, on average our PCA method actually decreased our performance when using cosine similarity. These results are shown in *table 3*.

K Value	Random PCA	Top 25 PCA
1	0.86609	0.8301
3	0.8301	0.82596
5	0.86609	0.81953
7	0.83286	0.82512

Table 3: Average Cosine Scores Across PCA Methods

Finally, there does not seem to be a consistent effect of k-value on performance. The average F1 measure data from *table 1* has been plotted in *figure 3* below.

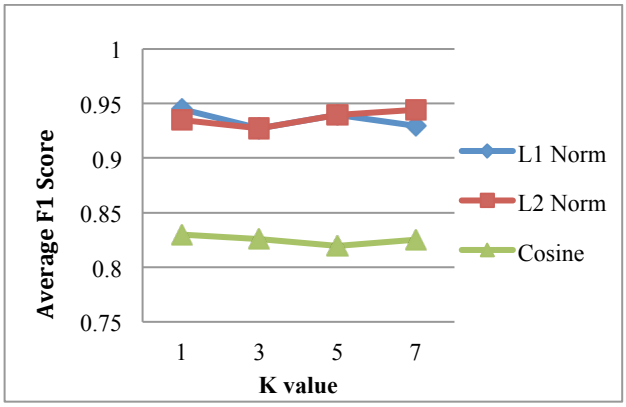

Figure 3: Average F score

No line consistently increases or decreases with k value. It does appear at the end that the L2 measure begins to increase with greater k value, though this is not consistent with performance at k = 3. Ultimately, we limited our highest k-value to 7 because of the limited size of our data.

6. CONCLUSION & FUTURE WORK

K-nearest neighbor methods perform well on the disease classification task. Somewhat unexpectedly, simpler proved to be better as our best performance came from using the L1 Norm with a k-value of 1. Cosine similarity performed the worst across all k values. Additionally, while we found that dimensionality reduction was necessary in order to allow efficient processing, we do not necessarily conclude that our method of PCA is the best. In fact, for certain measures of distance it may be harmful.

There are several possible reasons for these results. One possibility is that the signal to noise ratio needs to be further reduced. Further factor analysis may reveal a few specific genes that directly predict infection.

There did not seem to be a clear effect of k value. This may have to do with the relatively small size of the data set and its skew towards infected cells. Furthermore, as k increases past 7 the system would be more susceptible to noise among the uninfected cells and could over predict infection.

With a maximum accuracy of 92.075%, this classifier could be useful for medical applications and particularly useful as a part of a boosting approach, but is currently not high-enough to reliably diagnose patients. That said, our performance does suggest future research into this approach.

7. ACKNOWLEDGMENTS

We are thankful to Prof. Bryan Pardo for his lectures and his detailed feedback as we work on this project. We are also thankful to Prof. Rosemary Braun for her guidance on formulating the problem, finding data, and suggesting models for us to use. We also give our thanks to Prof. Doug Downey for his discussions on which model to use.

8. REFERENCES

[1] Erk, K., Padó, S. Exemplar-Based Models for Word Meaning in Context. 2010. *Proceedings of the ACL 2010 Conference Short Papers*. (July 11-16 2010). 92-97. http://dl.acm.org/citation.cfm?id=1858859

[2] Ioannidis, I., McNally, B., Willette, M., Peeples, M. E., Chaussabel, D., Durbin, J. E., Ramilo, O., Mejias, M., Flaño, E. 2012. Plasticity and Virus Specificity of the Airway Epithelial Cell Immune Response during Respiratory Virus Infection. *Journal of Virology*. 86(10):5422. DOI = 10.1128/JVI.06757-11.

[3] Pirooznia M.,Yang J, Qu M., Deng Y.2008. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics* 2008, 9(Suppl 1):S13. DOI = 10.1186/1471-2164-9-S1-S13

[4] Shapira SD, Gat-Viks I, Shum BO, Dricot A, de Grace MM, Wu L, Gupta PB, Hao T, Silver SJ, Root DE, Hill DE, Regev A, Hacohen N. 2009. A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell*, Dec 24, 139(7). DOI= 10.1016/j.cell.2009.12.018

[5] http://www.cdc.gov/rsv/research/us-surveillance.html

[6] http://www.cdc.gov/flu/pastseasons/1213season.html