Apr 5th, 8:00 AM - 12:00 PM

A Comparison of Unsupervised Methods for DNA Microarray Leukemia Data

Denise Harness

DEPARTMENT of MATHEMATICS & STATISTICS
College of Arts & Sciences
EAST TENNESSEE STATE UNIVERSITY

A Comparison of Unsupervised Methods for DNA Microarray Leukemia Data

Denise J. Harness | harnessd@etsu.edu | East Tennessee State University Honors College

Microarry Data

A DNA microarray measures the simultaneous expression of tens of thousands of genes in a patient. Rows are observations and columns are individual genes. The Leukemia data contains 72 observations and 7,129 genes. A subset of this data is shown below. Microarray data can be difficult to apply machine learning to because it typically has very few observations and many features.

	0	1	2	3	4	5	6	7	8	9
0	1.0	-1.462360	-0.645135	-0.835925	-1.470420	-0.919971	-1.584260	0.712393	-0.542291	1.050910
1	1.0	-0.664799	0.206146	-0.368575	0.258225	-0.475673	-0.354967	-1.119370	-0.292513	-0.375421
2	1.0	-0.200487	0.379941	-2.382780	0.439604	-1.226960	-1.762190	0.104636	-1.807490	0.492918
3	1.0	-0.257755	0.279937	1.839170	-1.629500	-1.287480	-1.265130	0.763342	-0.616454	-0.315784
4	1.0	-0.564569	-0.395885	-0.983716	-0.837410	-0.414772	0.148339	-0.035498	-0.100216	-0.757526

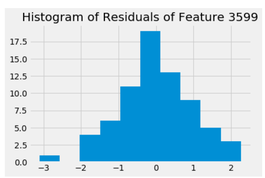
Multivariate Gaussian

The factors in a structured data set are often considered to be random variables jointly distributed as a multivariate Gaussian, as shown in Eq. 1:

$$f(X_1, X_2, ..., X_k) \sim \frac{1}{\sqrt{(2\pi)^k |C|}} e^{\frac{(k-\mu)^T C^{-1} (k-\mu)}{2}} \tag{1}$$

Where x is the vector of $X_1 ... X_n$. In practice, we assume a multivariate Gaussian distribution if empirical residuals are jointly Gaussian.

We view the residuals of the features and assume a Gaussian prior is present in the leukemia data.



References

This research was partially supported by NSF grant DUE-1356397.

[1] Jordi Belda, Luis Vergara, Addisson Salazar, and Gonzalo Safont. Estimating the laplacian matrix of gaussian mixtures for signal processing on graphs. *Signal Processing*, 2018.

[2] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

[3] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[4] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Covariance Matrix

The covariance of two random variables is the expected product of their deviations from their means. Covariance is reported in the product of the units of the variables. $C = \frac{1}{k-1} X^T X$ defines the empirical covariance matrix as a function of the leukemia data, X.

Precision Matrix

The inverse of the theoretical covariance matrix is simply the precision matrix, P, as shown in Eq. 2:

$$P = C^{-1} \tag{2}$$

The precision matrix is a way of showing associations or relationships among factors. This can be turned into a **probabilistic graphical model**, where each node in the graph is a gene with a probability associated to it. We need to **avoid** inverting the empirical covariance matrix unless it is both small and well-conditioned, which the Leukemia data is not.

Partial Correlation

It has been shown for the case of a Gaussian model there is an exact correspondence between the location of the non - zero entries in the precision matrix and the existence of partial correlations between the random variables. Partial correlations model the association between two features while adjusting for the effect of one or more additional features, as shown in Eq. 3:

$$\rho_{i,j} = \frac{-r_{i,j}}{\sqrt{r_{i,i} r_{j,j}}} \tag{3}$$

Partial correlations are ρ and the coefficients of the precision matrix are represented by $r_{i,j}$. Thus a probabilistic graphical model of the relationships (partial correlations) among features can be used as a surrogate for the true precision matrix.

Goal

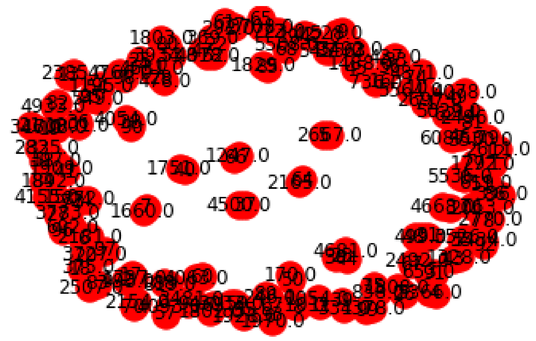
To compare the probabilistic modeling approach with traditional feature selection methods.

Methods

1. Standardize data and remove first column.
2. A Python 3 script was written to loop and create a matrix of partial correlations.
3. Lower triangular part of matrix and diagonal not calculated, because symmetric and all 1's.
4. Still have over 25 million entries, only top 100 partial correlation values and their corresponding positions in the matrix were stored.
5. The computation time on a single processor is about 10 days.
6. We ran 10% of the data at a time and saved after each run.
7. A network structure can be inferred using a Probabilistic Graphical Model, shown below in figure.
8. Reduced to 179 genes predicted to be the "more important" features in predicting Leukemia.

Results

The graph of the top 100 partial correlations is shown below. These are the features which make up the reduced feature set.



We used Random Forest and Linear Kernel Support Vector Machine algorithms to access the prediction accuracy of the reduced feature set. The mean and standard deviation of the accuracy of 100 iterations of the random forest algorithm is shown in the table.

	Full	**Reduced**
Mean	0.918	0.843
Standard Dev.	0.073	0.075

The prediction accuracies and F-Scores of the SVM algorithm were similar.

Conclusion

- Overall, the partial correlation method was able to reduce the full leukemia data set by 97%, from 7,129 features to 179 and maintain good prediction accuracy using both machine learning algorithms.
- These findings suggest the genes most responsible for Leukemia are among the 179 identified by the PGM approach.
- Although machine learning is sophisticated, it should not be performed independently from the field which the data came from. We would need to work with a biomedical researcher to confirm these findings.
- By knowing the $3 - 5\%$ most important genes in leukemia diagnosis, the activation of fewer genes must be studied, potentially leading to faster and less expensive results.
- This method can be applied to other data sets to learn about the underlying network structure between the features.

Future Research

Using more than 100 partial correlations could provide additional information about the gene activation in a patient with leukemia and increase the prediction accuracy. Future research may include considering other methods of producing more Laplacians matrices.