Gene Expression Analysis: Cancer Classification using Clustering

Kaushal Lodd, Saarang Rajguru

Abstract

We propose an unsupervised approach for cancer classification based on gene expression data, consisting of modifications to Minimum Spanning Tree (MST)-based clustering. We first apply the standard MST-based clustering algorithm to gene expression data from cancer patients and analyze its performance. We then propose two modifications to the algorithm to improve its accuracy in identifying cancer subtypes. The first modification involves the use of a distance metric that reduces standard deviation across the dataset. The second modification incorporates spectral clustering techniques in combination with MST methods to identify the genes that are most relevant for cancer classification. We evaluate the performance of the modified algorithms on publicly available cancer datasets and demonstrate that they outperform the standard MST-based clustering algorithm as well as other commonly used clustering methods, such as k-means. Our results suggest that the modified MST-based clustering algorithms have the potential to be useful tools for cancer diagnosis and classification.

Contents

1 Introduction

Gene expression analysis is a set of techniques used to measure the activity of genes in a cell or organism. The activity of a gene is usually determined by the level of the mRNA transcripts it produces, which is a reflection of the amount of protein it codes for. By measuring the mRNA levels, researchers can determine which genes are being actively expressed and at what levels, providing valuable insights into cellular processes and biological mechanisms. It is used in a variety of research fields, including genetics, developmental biology, and disease research. By understanding the patterns of gene expression, researchers can gain insights into how cells and organisms respond to various stimuli, and how diseases alter normal gene expression patterns.

It can help to identify different classes or subtypes of cancer based on their characteristic patterns of gene expression. By grouping together genes that have similar expression profiles, clustering algorithms can identify subgroups of cancer samples that are biologically distinct from each other. These subgroups can provide a way to classify cancers into different subtypes, based on their gene expression patterns.

In this paper, we focus on ALL (Acute Lymphoblastic Leukemia) and AML (Acute Myeloid Leukemia), which are two different types of leukemia, a cancer of the blood and bone marrow. The two types of leukemia differ in their clinical presentation, treatment,

and prognosis. ALL typically affects children and younger adults, and is generally considered more treatable than AML. AML, on the other hand, affects a broader age range and is generally considered more aggressive than ALL.

1.1 Dataset

We perform our analysis on a publicly available dataset published by Golub et al. in their paper.[1] There are two datasets containing the initial (training, 38 samples) and independent (test, 34 samples) datasets used in the paper. These datasets contain measurements corresponding to ALL and AML samples from Bone Marrow and Peripheral Blood. Intensity values have been re-scaled such that overall intensities for each chip are equivalent.

2 Proposed Methodology

2.1 Data Pre-processing

The original training dataset contains 7129 rows and 78 columns, the rows corresponding to the 7128 different genes that are considered, and the columns contain each of the 72 patients' gene expression levels, as well as gene description and gene accession numbers for each of the genes, apart from some other fields which are not relevant to our analysis.

We removed such columns, and performed a transpose operation on the dataset, such that we now have 72 patients as rows, and genes as columns.

In addition, we converted our dataset into pandas DataFrame format, for easier manipulation.

2.2 Metric: Silhouette Score

To evaluate the performance of our clustering techniques, we have chosen Silhouette Score[4]. It is a metric used to calculate the "goodness" of a clustering technique. Its value ranges from -1 to 1, where a value of -1 would mean that the clusters are incorrectly assigned, and a score of 1 would mean that the clusters are well apart from other and easily distinguished.

2.3 Euclidean MST (EMST)

A spanning tree is an acyclic subgraph of a graph G, which contains all the vertices from G. The minimum spanning tree (MST) of a weighted graph is the minimum-weight spanning tree of that graph. With the classical MST algorithms, the cost of constructing a minimum spanning tree is $O(m \log n)$, where m is the number of edges in the graph, and n is the number of vertices.

A Euclidean minimum spanning tree (EMST) is a spanning tree of a set of n points in a metric space (E^n), where the length of an edge is the Euclidean distance between a pair of points in the point set. The EMST clustering algorithm uses the Euclidean minimum spanning tree of a graph to produce the structure of point clusters in the n-dimensional Euclidean space.

Clustering based on EMST methods, which eliminates the longest edge by Euclidean distance, gave us a silhouette score of 0.194 and took 0.559s to execute.

2.4 Maximum Standard Deviation Reduction (MSDR)

First, an MST is generated from the input graph using a standard MST-based clustering algorithm. The standard deviation of the edge weights in the MST is then calculated, and

the edge with the highest weight is identified as the edge that contributes the most to the maximum standard deviation. The algorithm then attempts to reduce the maximum standard deviation by iteratively deleting edges from the MST. Edge deletion is performed based on a criterion that considers the impact of the edge on the maximum standard deviation of the MST. The process continues until no further reduction in the maximum standard deviation can be achieved. The resulting MST after edge deletion is the one with the lowest maximum standard deviation.[2]

This approach achieved a silhouette score of -0.088 and took 0.650s to execute.

2.5 Spectral Clustering Approach[3]

The following procedure is applied:

1. Constructed a similarity graph: Given a dataset, a similarity graph was constructed, where each data point represented a node in the graph and the edges between the nodes represented the similarity between the data points.

2. Computed the Laplacian matrix: The Laplacian matrix was computed from the similarity graph. The Laplacian matrix captured the graph structure and was used to define the spectral clustering algorithm.

3. Computed the eigenvectors of the Laplacian matrix: The eigenvectors of the Laplacian matrix were computed, and the k eigenvectors corresponding to the k smallest eigenvalues were selected.

4. Formed the matrix Y: The matrix Y was formed by stacking the k eigenvectors computed in the previous step.

5. Performed clustering: The MST and k-means algorithms were applied to the matrix Y to obtain the final clustering solution.

Spectral Clustering with k-means gave us a silhouette score of 0.746 and took 0.028s to execute, while spectral clustering with MST gave us a silhouette score of 0.697 and took 0.001s to execute.

3 Results & Discussion

3.1 Results

Method	Silhouette Score	Execution Time
EMST	0.194	0.559s
MSDR	-0.088	0.650s
Spectral k-means	**0.746**	0.028s
Spectral MST	0.697	**0.001s**

3.2 Discussion

From the results, it is clear that the spectral clustering techniques achieve significantly superior silhouette score as well as execution time.

Spectral MST takes the least execution time, and therefore it might be the most appropriate for resource-constrained settings.

However, best clustering performance is achieved by k-means clustering using spectral method.

References

[1] 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.

[2] Oleksandr Grygorash, Yan Zhou, and Zach Jorgensen. Minimum spanning tree based clustering algorithms. In *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*, pages 73–81. IEEE, 2006.

[3] R Jothi, Sraban Kumar Mohanty, and Aparajita Ojha. Functional grouping of similar genes using eigenanalysis on minimum spanning tree based neighborhood graph. *Computers in biology and medicine*, 71:135–148, 2016.

[4] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.