Enhanced Deterministic Clustering Methods for Cancer Subtype Prediction and Discovery from Genomic Data

Abstract

The last two decades witnessed massive advancements in the technology for the extraction of biomolecular data from tissue specimens, which resulted in the availability of high-quality genomic data. The adoption of Machine Learning methods to analyze such data has facilitated gaining more profound insights into the molecular basis of many diseases, particularly cancer. Cancer is a heterogeneous disease. A cancer type has multiple subtypes which differ from one another in the molecular events that trigger the disease. Machine Learning methods, especially clustering algorithms, have been instrumental in predicting and discovering molecular subtypes of cancers from genomic data. They provide better results than those of the conventional methods used for the task.

Various clustering methods which can be employed for the classification of cancer genomic data have been proposed in the literature. Despite this, the Biomedical research community has been preferring classical methods such as the *Hierarchical Clustering* method, for the task. Many methods lacked a considerable level of acceptance by the Biomedical research community due to the reasons such as the inherent complexity and bias, the need to assign values for a set of parameters for which finding an appropriate value is hard, and the non-deterministic nature. Accuracy is a critical aspect of 'cancer subtype prediction and discovery'. There is a pressing need for easy-to-use clustering methods having high accuracy and an ability to produce stable results. Moreover, cancer subtype discovery from genomic data is an exploratory analysis. It demands an automatic estimation of the number of *natural* clusters in the data which is a challenging task. Therefore, it is quite relevant to explore more effective and easy-to-use clustering methods that would facilitate 'cancer subtype prediction and discovery' from genomic data. The research work carried out in this direction has resulted in three significant contributions.

The first contribution is a density-based, deterministic variant of *K-Means* clustering algorithm for predicting the subtype of cancer from genomic data. *K-Means*

has been shown in the literature to be one of the most effective tools for classifying genomic data. The proposed method addresses two of the critical limitations of *K-Means*. Unlike the standard *K-Means*, the proposed method systematically finds the *initial centroids*. Moreover, it guarantees to find a group of well-separated data points as *initial centroids* which belong to dense regions in feature space.

The second contribution is a tool for cancer subtype discovery. The contribution is a Hierarchical Clustering algorithm which has a Silhouette Index-based objective function for selecting the pair of clusters to be merged in each step of the iterative process of building the clustering hierarchy. The method, unlike the other variants of the Hierarchical Clustering algorithm, estimates the number of *natural* clusters and finds the associated clustering solution.

The third contribution is also a tool for discovering cancer subtypes from genomic data. It is a partial ensemble clustering method which has two Hierarchical Clustering algorithms as components. The method estimates a set of the most likely numbers of *natural* clusters and finds the associated clustering solutions. The method leaves some data points unassigned to any cluster, and hence, called a partial clustering algorithm. Such unassigned data points are those for which there is no consensus between the component clustering algorithms, in the cluster assignments.

The performances of the proposed methods were compared with those of the state-of-the-art methods for cancer genomic data classification in terms of a prominent cluster validity index, namely *Adjusted Rand Index*. All the proposed methods were found to produce results with better accuracy when compared to the state-of-the-art methods. Besides, the results produced by the methods are stable, since the methods are deterministic. A group of ten cancer gene expression datasets, which includes some benchmark datasets, was used for the performance comparisons. Five general datasets were also used for the performance analysis to showcase the general applicability of the methods.

The proposed methods facilitate more accurate 'cancer subtype prediction and discovery'. Hence, these methods have considerable significance in contemporary cancer research. They can complement the efforts towards more effective implementation of the concept of 'Personalized Medicine'. The outcome of this research can also lead to the development of more specific and effective drugs, which would ultimately improve the survival rates of cancer patients.