A Novel Hybrid Dimensionality Reduction Technique for Cancer Classification from Microarray Data

Gobind Kumar Mandal[1], Dr. Manisha Kumari Deep[2]

[1]Research Scholar School of Computer Science and IT YBN University, Ranchi-834003, India
[2]Professor School of Computer Science and IT YBN University, Ranchi-834003, India

Abstract

Cancer classification using microarray data has become a critical area of research, given the complexity and high-dimensionality of genomic datasets. This paper proposes a novel hybrid dimensionality reduction technique for enhancing cancer classification accuracy by integrating two powerful methods: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). The proposed technique leverages PCA to capture the most significant global variance in the microarray data, followed by LDA to maximize class separability in the reduced feature space. The hybrid approach ensures that both global patterns and class-specific features are effectively preserved, improving classification performance. Experimental results on benchmark cancer microarray datasets demonstrate that the hybrid dimensionality reduction technique outperforms traditional methods, such as individual PCA and LDA, in terms of classification accuracy and computational efficiency. This method provides a promising solution to the challenges posed by high-dimensional genomic data, offering valuable insights for early cancer detection and personalized treatment strategies.

Keywords: Principal Component Analysis, DNA, PCA

INTRODUCTION

Cancer research has been transformed by the development of high-throughput gene expression profiling utilising microarray technology, which makes it possible to identify tumour subtypes and possible treatment targets with greater precision. In order to improve precision oncology, this study investigates the use of clustering and classification methods in the processing of cancer microarray data. Unsupervised clustering techniques, including k-means and hierarchical clustering, are essential for discovering new cancer subtypes since they classify tumours according to similarities in gene expression. However, by classifying samples into known classifications, supervised classification algorithms—such as support vector machines (SVMs), random forests, and deep learning models—help predict patient outcomes and treatment responses. The study draws attention to issues that can affect model performance, including high dimensionality, data noise, and class imbalance. Feature selection, dimensionality reduction, and ensemble learning techniques are examined as ways to deal with these problems. In order to improve the precision and dependability of classification algorithms, integration with clinical data and multi-omics techniques is also stressed. This study highlights the promise of clustering and classification in improving cancer diagnosis, prognosis, and individualised treatment plans by utilising cutting-edge computational approaches, opening the door to more successful precision oncology. Microarray technology's quick development has revolutionised cancer research by offering high-dimensional gene expression data that is essential for comprehending tumour heterogeneity. In order to support precision oncology, this research investigates the function of clustering and classification approaches in the analysis of cancer microarray data. Novel cancer subtypes can be found using unsupervised clustering techniques like k-means and hierarchical clustering, which are based on similarities in gene expression. Support vector machines (SVMs), decision trees, and deep learning models are examples of supervised classification algorithms that can identify cancer kinds, patient prognoses, and treatment outcomes. Notwithstanding their potential, these methods include drawbacks such class imbalance, noisy data, and excessive dimensionality. To improve model accuracy and dependability, techniques like feature selection, dimensionality reduction, and ensemble learning are used. Furthermore, combining clinical data and multi-omics data improves predictive models even more, leading to more individualised cancer treatment. This study emphasises the importance of clustering and classification techniques in enhancing cancer diagnosis and prognosis, which will ultimately aid in the creation of more accurate and potent oncology treatment plans. Precision oncology uses cutting-edge computational methods to evaluate enormous volumes of genomic data in order to customise cancer treatment based on unique molecular profiles. In order to determine tumour subtypes, forecast patient outcomes, and direct targeted therapy, this study investigates the function of clustering and classification techniques in cancer microarray data analysis.

Recently, a variety of DNA biochips with widgets and configurations have been developed. In addition to being profit-driven, these widgets allow DNA and/or RNA crossbreeding research to be carried out in the production of minuscule copies of highly parallel formats. Applications for DNA biochip crossbreeding are typically concentrated on transmission specimens for Sole Nucleotide Polymorphisms (SNPs) or DNA segment manifestation studies. Additionally, pharmacogenomics research, communicable diseases, cancer symptoms, criminal and inherited recognition goals, and molecule-based organically related studies and genomic research employing such biochip systems are being applied. The technology behind biochips continues to advance in terms of presentation in terms of discernment and kindness as well as in obtaining a more affordable study tool. DNA biochips will continue to revolutionise inherited research and other critical analytical fields. Furthermore, biochip technology created for DNA research is now being applied to new protein analysis directions in addition to living cell analysis.

Thousands of genetic variables can have their manifestation intensities traced simultaneously using a DNA biochip. Previous research has shown that this knowledge can help with cancer classification. Data on cancer biochips typically consists of a small number of specimens with a wide range of topographies and genetic factor manifestation intensities. Selecting relevant genetic variables involved in various cancer types remains a test. In addition to decreasing measurability, feature selection methods were carefully investigated to extract useful genetic component information from cancer biochip data.

Algorithm for K-Means Clustering

K is a user-defined variable that is used by the K-means clustering algorithm to find the K non-overlapping groups.

It initially selects K objects to serve as the centroids of the K clusters and assigns a cluster that is comparable to the one determined by the objects actual distance from the centroid to each of the remaining objects. After that, it calculates each of the new centroids and reassigns them based on their mistakes. The total of should ideally be used to apply the squared-error criterion (Rajoot et al. 2010). These centroids are further chosen in order to minimise the overall "error," where each error point is identified by a function that calculates the appropriate difference between a given point and its cluster centroid. This process is carried out again until no object changes and the criterion's function converges. The error provided by the cluster serves as a gauge of goodness. The K-means technique is used to calculate the squared error and the Euclidean distance, which is a gradient descent method for the minimisation of the squared error, gives.

The flowchart for the K Means centroid computation (Hartono & Abdullah 2015) is displayed in Figure.

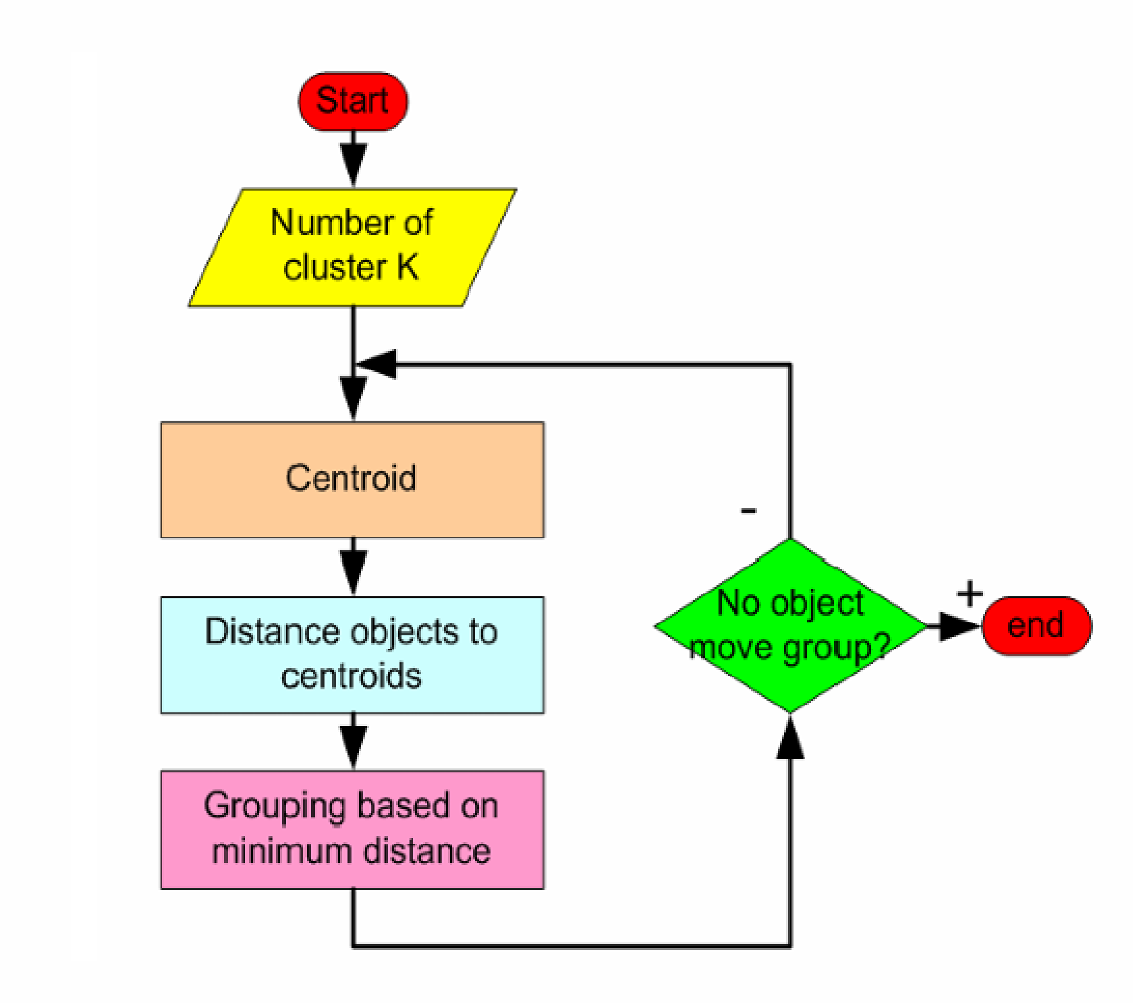
Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

Figure: K-Means Centroid computation mechanism

The steps of the k-means algorithm are listed below.

1. Initialization: To start the clusters, K input vectors or data points are selected.
2. The process of searching for nearest-neighbors involves identifying the closest cluster center for each input vector and assigning that input to the cluster that corresponds to it.

- Mean update: the mean or centroid for the input vectors assigned to the cluster is used to update the cluster centres in each centre.
- Stopping rule: until the means and their values no longer change, steps two and three are repeated.

There are some variations of the k-means that vary in how they choose the initial cluster centroids, the similarity metric, and the methods for calculating cluster means. The standard process for Euclidean data involves selecting initial centroids at random and using the mean as the centroid. Even though this is typically the local minimum, it converges to a single solution. Since only vectors are saved, O(m*n) of space is needed, where m is the number of points and n is the number of attributes. O(I*K*m*n) is the time requirement, where I is the number of iterations required for convergence.

Since the majority of the changes only occur in the first repetitions, the I is tiny and readily constrained. Accordingly, the K-means is a straightforward and efficient technique for clusters with less than m members.

In theory, this algorithm is thought of as a gradient descent technique that aims to minimise the sum of the squared errors of each point from the cluster centroid, or occasionally it is even thought of as a process that attempts to model the data.

Outcomes and conversation

The GEO Breast Cancer dataset and the Wisconsin Breast Cancer dataset were used in the experiments. The algorithms' classification accuracy, sensitivity, specificity, and f measure are assessed. Every experiment was conducted ten times.

Accuracy: the total number of classification characteristics (True Positive (TP) + True Negative (TN) + False Positive (FP) + False Negative (FN)) is divided by the accuracy, which is defined as the sum of the precise positive and negatives.

Accuracy= (TP+TN)/ (TP+TN+FP+FN)

Sensitivity, calculated as, is the percentage of positives that are accurately categorised as such.

Sensitivity=TP/(TP+FN)

Specificity, as described by, is the percentage of negatives that are accurately categorised as such.

Specificity=TN/(TN+FP)

The precision and recall harmonic means in, or F-measure, are the F-measure.

F-Measure = (2*(Precision*Recall))/(Precision+Recall)

Where

Precision=TP/ (TP+FP)

Recall=TP/ (TP+FN)

Misclassification rate=1 - (TP+TN)/(TP+TN+FP+FN)

All of the aforementioned evaluation metrics, which are provided from Table for the comparative analysis of mutual information-based supervised gene clustering, have experimental observations recorded. (MSG), conventional K Means clustering and Sparse Probabilistic Principal Component Analysis (SPPCA). Two widely used, publicly accessible standard datasets for breast cancer were used in all of the experiments: the Gene Expression Omnibus Database, which contains 738 occurrences and the Wisconsin Breast Cancer Dataset, which contains 699 instances. Every experiment was conducted ten times. The accuracy, sensitivity, specificity, F-measure and misclassification rate of K-Means Clustering have all been examined.

Conclusions

In order to recover the cluster value, K-Means clustering is used in this section of the work to find missing values in the microarray gene appearance data. These methods are employed to obtain the fewest possible unforeseen gene data sets, and in the past, it improved the groups' quality and accuracy by using the K-Means gathering technique. The following is a summary of the observations drawn from the results obtained.

- In the Wisconsin dataset, K Means clustering increased classification accuracy by 2.24% compared to SPPCA and 4.03% compared to MSG. In the GEO dataset, K Means clustering increased classification accuracy by 0.47% compared to SPPCA and 1.80% compared to MSG.
- In the Wisconsin dataset, the K Means clustering increased specificity by 3.88% compared to SPPCA and 5.88% compared to MSG. In the GEO dataset, the K Means clustering increased specificity by 0.59% compared to SPPCA and 1.20% compared to MSG.
- In the Wisconsin dataset, K Means clustering increased sensitivity by 1.43% compared to SPPCA and 3.13% compared to MSG, whereas in the GEO dataset, K Means clustering increased sensitivity by 0.4% compared to SPPCA and 4.25% compared to MSG.
- In the Wisconsin dataset, K Means clustering has improved F-Measure by 3.47% compared to SPPCA and 6.09% compared to MSG, whereas in the GEO dataset, it has increased F-Measure by 0.80% compared to SPPCA and 3.64% compared to MSG.

The K-Means algorithm's arbitrary selection of the initial seed set of favoured clusters is one of its drawbacks. Additionally, because it is presumed that each predefined feature has the same weight; it is quite difficult to determine which attribute contributes more to the grouping process.

References

1. Han, F, Sun, W & Ling, QH 2014, ‗A novel strategy for gene selection of microarray data based on gene-to-class sensitivity information', PloS one, 9(5), e97530. https://doi.org/ 10.1371/ journal. pone.0097530.
2. Han, F, Yang, C, Wu, YQ, Zhu, JS, Ling, QH, Song, Q & Huang, D 2015, ‗A Gene Selection Method for Microarray Data Based on Binary PSO Encoding Gene-to-class Sensitivity Information', IEEE/ACM Transactions on Computational Biology and Bioinformatics(TCBB), vol. 14, no. 1, pp. 85-96.
3. Hartigan, John A & Manchek A Wong 1979, ‗Algorithm AS 136: A k means clustering algorithm', Journal of the Royal Statistical Society, Series C (Applied Statistics) vol. 28, no. 1, pp. 100-108.
4. Hartono, EO & Abdullah, D 2015, ‗Determining a Cluster Centroid of K-Means Clustering Using Genetic Algorithm'. International Journal of Computer Science and Software Engineering (IJCSSE), vol. 4, no.
5. Hassanzadeh, T &Meybodi, MR 2012, ‗A new hybrid approach for data clustering using firefly algorithm and K-means', Artificial Intelligence and Signal Processing (AISP), 16th CSI International Symposium on IEEE, vol. 1, pp. 007-011.
6. Heller & Michael, J 2002, ‗DNA microarray technology: devices, systems, and applications', Annual review of biomedical engineering 4, vol. 4, no. 1, pp. 129-153.
7. Hsu, HH, Hsieh, CW & Lu, MD 2011, ‗Hybrid feature selection by combining filters and wrappers'. Expert Systems with Applications, vol. 38, no. 7, pp. 8144-8150.
8. Hu, H, Li, J, Wang, H & Daggard, G 2006, ‗Combined gene selection methods for microarray data analysis', Springer Berlin Heidelberg International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, vol. 1, pp. 976-983.
9. Hussain, SF & Ramazan, M 2016, ‗Biclustering of human cancer microarray data using co-

similarity based co-clustering', Expert Systems with Applications, vol. 55, pp. 520-531.

10. Jiang, D, Tang, C & Zhang, A 2004, _Cluster analysis for gene expression data: A survey', IEEE Transactions on knowledge and data engineering, vol. 16, no. 11, pp. 1370-1386.
11. Jin, C & Jin, SW 2016, _Gene selection approach based on improved swarm intelligent optimisation algorithm tumour classification', IET systems biology, vol. 10, no. 3, pp. 107-115.
12. Jin, X, Xu, A, Bie, R & Guo, P 2006, _Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles'. In International Workshop on Data Mining for Biomedical Applications, Springer, Berlin, Heidelberg, pp. 106 115.
13. Kar, S, Sharma, KD & Maitra, M 2015, _Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique', Expert Systems with Applications, vol. 42, no. 1, pp. 612-627.
14. Kaur, J & Gurm, JS 2015, _Optimizing the Accuracy of CART Algorithm by Using Genetic Algorithm', International Journal of Computer Science Trends and Technology (IJCST), vol. 3 no. 4, pp. 142-147.
15. Kaur, S, Sharma, AS, Kaur, H & Singh, K 2016, _Gene selection for tumor classification using resilient backpropagation neural network', International Conference on Advances in Computing, Communication, & Automation (ICACCA)(Fall), vol. 1, pp. 1-5.
16. Keerin, P, Kurutach, W & Boongoen, T 2016, _A cluster-directed framework for neighbour based imputation of missing value in microarray data', International Journal of Data Mining and Bioinformatics, vol. 15, no. 2, pp. 165-193.