International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Data Mining Techniques in Bioinformatics Analysis

C. Kondal Raj[1], Dr. R. Murugesan[2]

Submitted: 16/05/2024 **Revised**: 29/06/2024 **Accepted**: 09/07/2024

Abstract: Microarray experiments yield vast datasets containing expression data for thousands of genes across a limited number of samples, usually no more than a few dozen. A major challenge is identifying groups of genes that are co-regulated and collectively show strong associations with specific outcome variables. To tackle this challenge, we suggest using k-means clustering algorithms, which leverage external information about response variables to group genes effectively. We propose an algorithm based on logistic regression analysis that integrates gene selection, supervision, gene clustering, and sample classification into a single streamlined process. Through empirical studies on diverse microarray datasets, we demonstrate its ability to pinpoint gene clusters whose expression centroids exhibit robust predictive potential, surpassing conventional methods focused on individual gene analysis. This approach not only promises advancements in medical diagnostics and prognostics but also enhances functional genomics by offering insights into gene function and regulation.

Keywords: *Data Mining, Bioinformatics, Microarray Datasets, k-means*

1. Introduction

Microarray technology offers a powerful method for monitoring gene expression on a large scale, holding great promise for advancing medical diagnostics and functional genomics. With robust statistical techniques tailored for analyzing extensive gene expression datasets, we aim to achieve precise classification of tumor subtypes. This capability could pave the way for personalized treatments that optimize effectiveness while minimizing side effects. Gene expression data also play a crucial role in reconstructing gene regulatory networks, offering deeper insights into genome functionality. A significant challenge lies in identifying groups of genes that work together, such as within pathways, whose combined expression effectively predicts specific outcomes (y). Our objective is to establish rules like: "high centroids of gene 534, gene 837, and gene 235, coupled with low centroids of gene 2194, gene 1438, gene 931, and gene 694, indicate cancer subtype A." These gene groups and their centroids could serve as markers for accurately predicting phenotypes in medical diagnostics and enhancing understanding of biological processes and gene regulation. However, this task is complex due to computational challenges posed by the large number of predictor variables (genes) and statistical issues arising from the "small n, large p" problem.

In the pursuit of identifying co-regulated genes, researchers commonly employ unsupervised clustering algorithms such as hierarchical clustering, kmeans clustering, self-organizing maps, and principal components analysis. These methods group genes based on similarity measures derived solely from their expression profiles, without considering variations in the y-values (response variables). However, our objective is to uncover groups of co-regulated genes strongly associated with the response variable y. To achieve this, we propose the use of supervised clustering algorithms, which integrate predictor variables controlled by x-values with external supervised information derived from yvalues.

Previous research has explored techniques like partial least squares, traditionally used in chemometrics, which constructs weighted linear combinations of genes that exhibit maximal covariance with the outcome variable. One limitation of this approach is that each fitted component involves all genes, often numbering in the thousands, rather than identifying smaller, biologically meaningful clusters of genes acting in concert within pathways. Consequently, while partial least squares provides statistical relationships, it lacks the biological insight derived from identifying clusters of genes that function similarly.

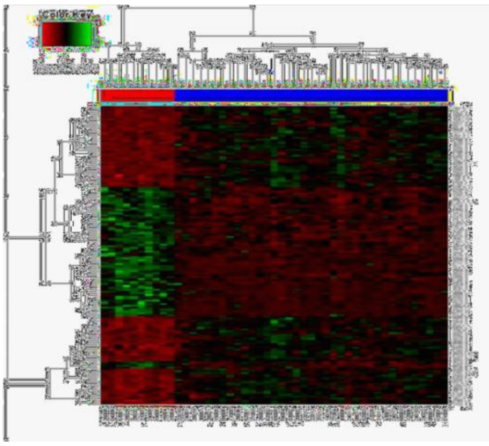

Fig 1. Micro array data

[1] *Assistant Professor, Department of Computer Science and Information Technology, CPA College (Affiliated to Madurai Kamaraj University), Bodinayakanur, Theni, Tamil Nadu, India*

[2] *Associate Professor, Department of Computer Science, CPA College (Affiliated to Madurai Kamaraj University), Bodinayakanur, Theni, Tamil Nadu, India*

A different approach to supervised clustering that addresses these limitations is tree harvesting, a method that involves two main steps. Initially, it generates numerous candidate clusters using unsupervised hierarchical clustering. Subsequently, all cluster centroids are considered as potential input variables in a supervised model aimed at discriminating tissue types, with the selection of gene groups that offer the most discriminatory value. However, this method fixes the clustering once it's determined.

A more integrated strategy involves combining gene selection, clustering, and supervision into a single cohesive step. This approach was first introduced in our work on "Supervised Clustering of Genes". Subsequently, similar single-step methods were also explored by Jornsten and Yu. Our generic strategy for supervised clustering focuses on sequentially optimizing an empirical objective function that measures how effectively clusters explain the outcome variable Y.

Initially, our implementation, named Wilma, used clustering criteria based on Wilcoxon and margin statistics. Subsequently, we developed a novel approach to k-means clustering, leveraging an objective function rooted in logistic regression analysis. This approach enhances upon Wilma in several ways: it can capture genes involved in multiple pathways, fosters stronger interactions between clusters, demonstrates greater robustness, allows for the inclusion of additional clinical covariates to refine clustering, and can be easily adapted to continuous response problems. Moreover, it incorporates a built-in classifier, resulting not only in theoretical and methodological advancements but also in outstanding empirical prediction performance.

2. Related Work

Researchers from various disciplines are increasingly focusing on microarray data analysis. The development of classification methods for these datasets has been a significant area of interest. Recent advancements in artificial intelligence, machine learning, pattern recognition, and related fields have led to the proposal of several innovative approaches. These methods leverage the progress in biomedical and information technologies to explore diverse algorithms aimed at enhancing cancer diagnosis through data-driven diagnostic techniques.

Mabu *et al*. proposed a method combining cluster-based feature selection and artificial neural networks for classifying gene expression datasets. In a similar vein, Zeebaree et al. introduced an approach using gene selection and convolutional neural networks for microarray cancer data classification. However, they did not detail how they reduced the dimensionality of the original datasets.

Hou *et al.* presented a diagnostic prediction model for prostate cancer, integrating an optimized genetic algorithm with artificial neural networks. They validated their model using prostate cancer datasets.

Mohapatra *et al.* suggested using ridge regression (RR) combined with a single hidden layer feed-forward network, where feature weights were randomly assigned. Their validation involved binary microarray datasets such as Breast, Prostate, Colon tumor, and Leukemia. It was noted that the standard train/test protocol was not consistently applied, particularly in the case of the breast cancer dataset.

Salem *et al.* put forward a genetic programming-based cancer classifier utilizing information gain (IG) for feature selection. In a different approach, Lin *et al.* employed a genetic algorithm combined with silhouette statistics to select features and classify the SRBCT dataset. However, there were concerns about the nonoptimality of their feature selection method, which generated numerous features leading to over-fitting.

Sharbaf *et al.* proposed a hybrid method for gene selection and

classification of microarray datasets using cellular learning automata and ant colony optimization. They explored the impact of various rank-based feature selection techniques and validated their approach using three classifiers: support vector machine (SVM), k-nearest neighbor (KNN), and Naive Bayes.

Kumar *et al.* developed algorithms for feature selection and classification using the MapReduce framework in combination with the K-nearest neighbors (KNN) classifier. Their approach aimed to optimize the processing of large-scale datasets through distributed computing. Nguyen et al. introduced an aggregate gene selection method for classifying microarray data, evaluating their model across four standard datasets: DLBCL, Leukemia, Prostate, and Colon. They validated their approach using five different classifiers: linear discriminant analysis, KNN, probabilistic neural network, SVM, and multilayer perceptron (MLP). They claimed stability across these classifiers but did not extend their evaluation beyond these five.

Lofti and Keshavarz proposed a hybrid approach combining Principal Component Analysis (PCA) with brain emotional learning for classifying microarray cancer data. Their validation covered three datasets, but the scope may not fully confirm the method's general applicability. Ravi et al. conducted a comprehensive review highlighting the potential of deep learning models in health informatics. They discussed various architectures such as deep feed-forward, convolutional networks, and recurrent networks applicable to diverse problem domains.

Kar *et al*. proposed a particle swarm optimization-based feature selection method for classifying microarray cancer data, validating their approach on datasets like ALL-AML and SRBCT through multiple experimental runs. Garcia and Sanchez suggested a two-stage method for microarray classification. They employed the ReliefF ranking algorithm for feature selection and evaluated their approach using three linear classifiers: Fisher linear discriminant, SVM, and MLP neural network, reporting results across eight cancer datasets.

Chen *et al.* applied particle swarm optimization-based feature selection with the C4.5 decision tree for classifying tumor cancer datasets, employing a 5-fold cross-validation approach to assess their method's performance. Farid *et al.* proposed an adaptive rule-based classifier for large biological datasets, incorporating decision tree and KNN techniques, although it lacked adaptability concerning the number of neighbors.

Lyu *et al*. introduced a filter-based feature selection method using maximum information coefficient and Gram-Schmidt orthogonalization for microarray cancer data classification. Li *et al*. devised an overlapped grouping strategy with data-driven weights based on information theory for lung cancer classification. Piao *et al.* introduced a feature subset-based ensemble method to classify multi-class microarray cancer data, learning from different projections of the original feature space. Wang *et al.* proposed an integrated Markov blanket technique and Wrapper-based feature selection method to handle computational complexity caused by redundant features during feature selection. Hoque *et al.* introduced a greedy feature selection technique utilizing mutual information for feature-feature and feature-class relationships, validated across three base classifiers: KNN, Random Forest (RF), and SVM.

3. Proposed Methods

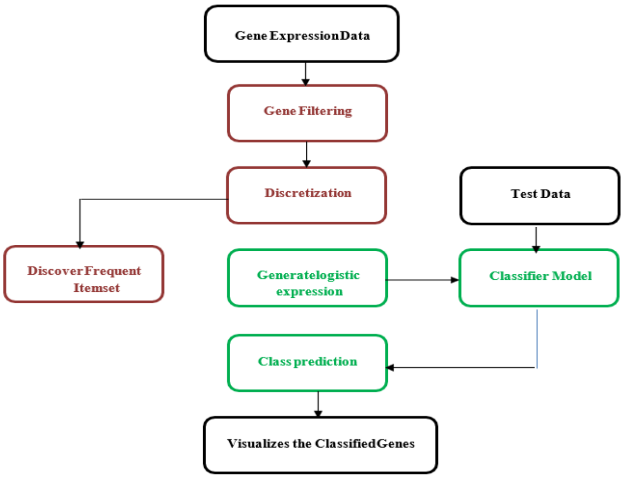

Fig 2 Proposed method

3.1 K-MEANS CLUSTERING

When clustering genes involved in various stages of erythropoiesis, unsupervised learning is necessary because the relationships between genes and cell development stages are not fully understood. Among the representative-based clustering methods, k-means stands out. It divides a dataset D $= \{x_i\}_{i=1}^{n}$ into k clusters denoted as C={C1,C2,…,Ck}. Each cluster is characterized by a representative point, typically chosen as the centroid (µ) of all points within that cluster.

These algorithms depend on assessing the similarity among observations within clusters. K-means is among the most widely used and straightforward clustering methods. It employs an iterative approach aimed at minimizing the total squared distance across all observations within clusters.

Logistic Regression (LR) is a well-known statistical technique used for binary classification, particularly in modeling binary data. Let d represent a vector of independent or feature variables, and let {-1, +1} denote the respective binary class labels. The logistic model can be defined as

y/x = 1 1 + exp(-y(β Tx + α)) = exp(y (β Tx + α)) 1+ exp(y (β Tx + α)) (1)

The expression $r(y/X)$ denotes the conditional probability of y given the features x. In the logistic model, the parameters consist of α € R and β €R where α, β includes the intercept term and the weight vector term.

βTx + α = 0 defines a hyperplane within $p(y/X)$ = 0.5.

If the conditional probability $p(y/x)$larger than 0.5, Tx + α = 0

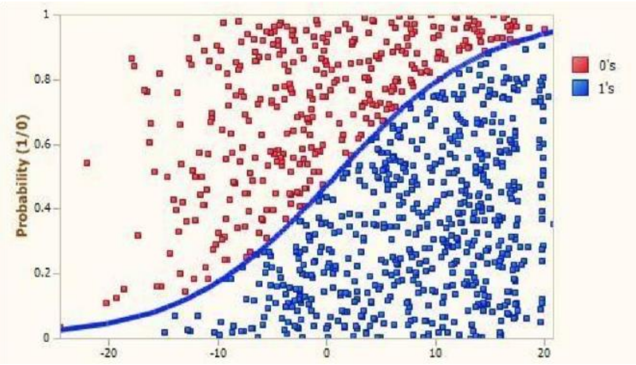

Fig 4 Logistic Regression

Newton method will be used in this paper. After obtaining the maximum likelihood values of and which are the solutions of (7), finally the probability of the two possible outcomes will be predicted entering by entering a new features vector x, to the associated logistic regression model; the logistic regression classifier is formed as: ( ) = $sg(\beta Tx +)$ (8)

Where $sgn( ) = \{ +1 > 0 -1 \leq 0.$

4. Numerical Results

We assessed our clustering algorithms using various datasets that detail gene expression among cancer patients. Specifically, we examined: The leukemia dataset, which includes gene expression data from n = 72 patients diagnosed with either acute lymphoblastic leukemia (ALL, 47 cases) or acute myeloid leukemia (AML, 25 cases). This dataset was derived from Affymetrix oligonucleotide microarrays and comprised a training set of 38 observations and a test set of 34 samples. Prior to analysis, we pre-processed the data through thresholding, filtering, base 10 log-transformation, and standardization. Ultimately, the dataset encompassed expression values from p = 3,571 genes. We examined multiple datasets tracking gene expression in various cancer types:

The estrogen and nodal datasets monitor p = 7,129 genes across 49 breast tumor samples using Affymetrix technology. After applying thresholding between 100 and 16,000 expression units, we performed base 10 logtransformation and standardized each experiment to zero mean and unit variance. These datasets include response variables indicating estrogen receptor status and lymph node involvement. The colon cancer dataset, also obtained via Affymetrix technology, features expression levels from 40 tumor and 22 normal colon tissues for a selected set of 2,000 genes with high minimal intensity across samples. We further processed this data with base 10 log-transformation and standardized each experiment to zero mean and unit variance.

The prostate cancer dataset comprises expression data from 52 prostate tumor and 50 non-tumor samples, obtained using Affymetrix technology. We utilized normalized and thresholded data, resulting in base 10 logtransformed expression values for p = 6,033 genes, standardized to zero mean and unit variance across genes. The lymphoma dataset involves cDNA microarray gene expression across the three most prevalent adult lymphoid malignancies (K = 3). The dataset includes n = 62 samples and documents the expression of 4,026 genes known for their relevance in lymphoid cells or immunological and oncological contexts. We addressed missing values through imputation and standardized the entire dataset.

5. Conclusion

We have introduced a methodology for k-means clustering of genes derived from microarray experiments, which holds potential benefits for medical diagnostics and prognostics. This approach identifies clusters of genes that interact closely, with centroids of their expressions exhibiting strong explanatory power for the response variable. These gene clusters and their centroids enable accurate prediction of outcomes for new samples. However, k-means clustering serves not only as a predictive tool but also as a foundational method for exploring gene function, regulation, and genome dynamics. Our novel supervised clustering algorithm, logistic regression, integrates gene selection, supervision, gene clustering, and optional sample classification into a unified approach. It aims to identify gene clusters whose centroids facilitate straightforward discrimination of the outcome variable

y. This method constructs clusters incrementally through a combination of forward steps and periodic refinement steps, guided by an empirical objective function incorporating information from y-values and conditional class probabilities derived from penalized logistic regression analysis.

Logistic regression overcomes several limitations of traditional k-means clustering by capturing genes involved in multiple pathways without requiring cluster disjointness. By adopting a criterion based on multiple clusters, it identifies cohesive groups of interacting genes rather than treating each gene individually. Additionally, we propose extensions of logistic regression to handle polytomous and continuous response problems, as well as its integration with clinical covariates. Beyond its robust features and cohesive algorithm grounded in rigorous statistical methodology, logistic regression is supported by comprehensive empirical validation across diverse microarray gene expression datasets. Our findings provide compelling evidence of its practical efficacy and suitability for real-world applications. The logistic regression built-in classifier not only outperforms other methods but also surpasses established classifiers and advanced machine learning techniques that operate with individual genes as inputs. While initially designed for microarray data analysis, the combination of logistic regression and k-means clustering shows promise for other datasets facing the challenge of "large p, small n," where a few key clusters are likely to significantly influence outcome variations.

References

[1] Nguyen D, Rocke D: Tumor Classification by Partial Least Squares Using Microarray Gene Expression Data. Bioinformatics 2002, 18: 39–50.

[2] Hastie T, Tibshirani R, Botstein D, Brown P: Supervised Harvesting of Expression Trees. Genome Biology 2001, 1: 1–12.

[3] Dettling M, B¨uhlmann P: Supervised Clustering of Genes. Genome Biology 2002, 3: research 0069.1–0069.15.

[4] J¨ornsten R, Yu B. Simultaneous Gene Clustering and Subset Selection for Sample Classification via MDL. To appear in Bioinformatics 2003.

[5] Bickel P, Klaassen C, Ritov Y, Wellner J: Efficient and Adaptive Estimation for Semiparametric Models. John Hopkins University Press, 1993.

[6] Dudoit S, Fridlyand J: A Prediction-Based Resampling Method to Estimate the Number of Clusters in a Dataset. Genome Biology 2002, 3(7): 0036.1–0036.21.

[7] Tibshirani R, Walther G, Hastie T: Estimating the Number of Clusters in a Dataset via the Gap Statistic. Technical Report 208, Department of Statistics, Stanford University, 2000.

[8] La Cessie S, Van Houwelingen J: Ridge Estimators in Logistic Regression. Applied Statistics 1990, 41, 191–201.

[9] Eilers P, Boer J, Van Ommen G, Van Houwelingen H: Classification of Microarray Data with Penalized Logistic Regression. Proceedings of SPIE 2001, Volume 4266: Progress in biomedical optics and imaging, 2: 187–198.

[10] Zhu J, Hastie T: Classification of Gene Microarrays by Penalized Logistic Regression. Preprint, Department of Statistics, Stanford University, 2002.

[11] Dettling M, B¨uhlmann P: Boosting for Tumor Classification with Gene Expression Data. To appear in Bioinformatics 2003.

[12] Allwein E, Schapire R, Singer Y: Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. Journal of Machine Learning Research 2000, 1: 113–141.

[13] Hoerl A, Kennard R: Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 1970, 12: 55–67.

[14] Golub T, Slonim D, Tamayo P, Huard C, Gassenbeek M, Coller H, Loh M, Downing J, Caliguri M, Bloomfield C, Lander E: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 1999, 286: 531–538

[15] Dudoit S, Fridlyand J, Speed T: Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of the American Statistical Association 2002, 97: 77–87.