INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Cancer Gene Identification through Clustering

V. Sridevi
Assistant Professor, Department of Computer Applications
Dr. N.G.P.Arts and Science College, Coimbatore, Tamil Nadu, India
P. Vidhya
Research Scholar, Department of Computer Science,
Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India

Abstract:
Cancer is one of the dangerous diseases, which is more severe to cure. Sometimes, certain types of cancer seem to run in some families. Gene prediction is an emerging research area that had received growing attention in the research community. All the genes may not be biologically significant in diagnosing the disease. Microarray cancer data organized as samples versus genes fashion are being exploited for the classification of tissue samples into benign and malignant or their subtypes. This paper surveys various data mining methods involved in cancer gene identification.

1. Introduction

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Knowledge discovery is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools is used to predict the behaviors and future trends. It allows business to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. Data Mining techniques are efficiently used to get similarities between searching for valuable information in a large database.

2. Data Mining Techniques

2.1. Classification

Classification is one of the powerful data mining techniques, which is used to classify the huge amount of data. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification.

2.2. Clustering

Clustering can be defined as identification of similar classes of objects. By using clustering techniques it is possible to identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes.

3. Prediction

Regression technique can be adapted for prediction. Regression analysis is commonly used to describe the model that shows the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are to predict the need. Unfortunately, many real-world problems are not simply prediction. The more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values.

4. Association Rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one.

5. Introduction to Gene Expression

With the recent advancement of DNA microarray technologies, the expression levels of thousands of genes can be measured simultaneously. The obtained data are usually organized as a matrix (also known as a gene expression profile), which consists of n columns and m rows. The columns represent genes (usually genes of the whole genome), and the rows correspond to the samples (e.g. various tissues, experimental conditions, or time points).Given this rich amount of gene expression data, the goal of microarray analysis is to extract hidden knowledge (e.g., similarity or dependency between genes) from this matrix. The analysis of gene expression may identify mechanisms of gene regulation and interaction, which can be used to understand a function of a cell.One of the key steps in gene expression analysis is to perform clustering genes that show similar patterns. By identifying a set of gene clusters, we can hypothesize that the genes clustered together tend to be functionally related. With the abundance of microarray data, genome-wide expression data clustering has received Significant attention during the past few years in the bioinformatics research community, ranging from hierarchical clustering, self organizing maps, and neural networks, algorithms based on Principal Components Analysis or Singular Value Decomposition subspace clustering, and graph-based approach.

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as ribosomal RNA (rRNA), transfer RNA (tRNA) or small nuclear RNA (snRNA) genes, the product is a functional RNA. The process of gene expression is used by all known life - eukaryotes (including multicellular organisms), prokaryotes (bacteria and archaea), possibly induced by viruses - to generate the macromolecular machinery for life.

Several steps in the gene expression process may be modulated, including the transcription, RNA splicing, translation, and post-translational modification of a protein. Gene regulation gives the cell control over structure and function, and is the basis for cellular differentiation, morphogenesis and the versatility and adaptability of any organism.

6. Gene Markers

A genetic marker is a gene or DNA sequence with a known location on a chromosome that can be used to identify individuals or species. It can be described as a variation (which may arise due to mutation or alteration in the genomic loci) that can be observed. Genetic markers can be used to study the relationship between an inherited disease and its genetic cause. It is known that pieces of DNA that lie near each other on a chromosome tend to be inherited together. This property enables the use of a marker, which can then be used to determine the precise inheritance pattern of the gene that has not yet been exactly localized.

7. General Survey

Wang, et al., [1] demonstrates that DNA micro array can pursue the expressions of many genes simultaneously. Micro-array data habitually a surround a petite number of samples, it includes a hefty number of gene expression levels as a feature. It is a challenging task to choose relevant genes involved in different types of cancer. For the purpose of mining information about genes from a cancer micro-array data and dimensionality reduction, the algorithm such as feature selection algorithms was systematically analyzed.

F. Chu and L. Wang [2] stated that Micro array gene expression data generally have a huge number of dimensions. The classifier used here is a support vector machine (SVM for cancer classification with the microarray gene expression data. The selection of genes has been completed by the use of four effective feature dimensionality reduction methods, for instance, principal components analysis (PCA), class- separability measure, Fisher ratio, and T-test. The data set used here is SRBCT, lymphoma data set and leukemia data set of publicly available micro array gene expression data set.

In [8], Clinical responses to anticancer therapies are often restricted to a subset of patients. In some cases, mutated cancer genes are potent biomarkers for responses to targeted agents. Here, to uncover new biomarkers of sensitivity and resistance to cancer therapeutics, we screened a panel of several hundred cancer cell lines which represent much of the tissue-type and genetic diversity of human cancers with 130 drugs under clinical and preclinical investigation. In aggregate, we found that mutated cancer genes were associated with cellular response to most currently available cancer drugs.

Huilin Xiong and Xue-Wen Chen [3] says the new approach called kernel function, which improves the performance of the classifier in genetic data. The efficiency of a kernel approach has been probed in which it is depends upon on optimizing a data-dependent kernel model. The K-nearest-neighbor (KNN) and support vector machine (SVM) could be used as a classifier for performance analysis. Data set utilized here, ALL-AML Leukemia Data, Breast-ER, Breast-LN,, Colon Tumor Data, Lung Cancer Data and Prostate Cancer from micro array data.

L. Shen And E.C. Tan [4] presented the penalized logistic regression for classification of cancer. The penalized logistic regression united with two-dimension reduction methods in order that the classification accuracy and computational speed were improved. Support vector machines and least squares regression chose for comparison. The method called the Recursive feature elimination (RFE) was used for iterative gene selection, which tries to select a gene subset that was most relevant to the cancers. Seven publicly available data sets such as breast cancer, central nervous system, colon tumor, Acute Leukemia, Lung cancer, ovarian cancer and Prostate cancer data set were chosen from [8] to performance evaluation. Linear SVM is used to compare with the regression methods.

Mathew J. Garnett et. al proposed a new clustering algorithm that are applied to the analysis of gene expression data are converted into a distance matrix, a weighted graph is constructed according to the combined matrix, and a graph partitioning approach which is used to cluster the graph to generate the final clusters. Consensus clustering obtained from clustering multiple times with Variational Bayes mixtures of Gaussians have been successfully applied to the unsupervised analysis of functional classes of genes in yeast [7],

while a graph-based ensemble clustering algorithm has been recently proposed to discover the underlying classes of the examples in gene expression data.

8. Conclusion

Gene prediction is a rising research area that has received growing attention in the research community over the past decades. The cancer classification using gene expression data is known to contain the keys for addressing the fundamental problems related to cancer diagnosis. This paper analyzed various research papers that proposed different techniques to identify the cancer gene. Clustering is one of the important methods which is already used for cancer gene identification. From this survey we conclude that the Support Vector Machine is efficiently working in field of cancer gene identification.

9. References

1. Wang,Y.,Tetko, I. -V., Hall, M. -A., Frank, E., Facius, A., Mayer, K. -F., And Mewes H. -W., "Gene Selection From Microarray Data For Cancer Classification —A Machine Learning Approach", Comput Biol Chem, 29 (1): 37-46, 2005.
2. F. Chu and L. Wang, "Applications Of Support Vector Machines To Cancer Classification With Microarray Data", International Journal Of Neural Systems, Vol. 15, No. 6, 475–484, 2005.
3. Huilin Xiong And Xue-Wen Chen,"Optimized Kernel Machines For Cancer Classification Using Gene Expression Data", Proceedings Of The 2005 IEEE Symposium On Computational Intelligence In Bioinformatics And Computational Biology, Pp.1-7, 2005.
4. L. Shen And E.C. Tan, "Dimension Reduction-Based Penalized Logistic Regression For Cancer Classification Using Microarray Data," IEEE/ACM Trans. Computational Biology And Bioinformatics, Vol. 2, No. 2, Pp. 166-175, Apr.-June 2005.
5. E. Cox, Fuzzy Modeling And Genetic Algorithms For Data Mining And Exploration, Elsevier, 2005
6. J. C. Bezdek, Fuzzy Mathematics in Pattern Classification, Ph.D. thesis, Center for Applied Mathematics, Cornell University, Ithica, N.Y., 1973.
7. G. J Klir, T A. Folger, Fuzzy Sets, Uncertainty and Information, Prentice Hall,1988
8. Mathew J. Garnett, Patricia Greninger, I. Richard Thompson, Xi Luo, Jorge Soares, Qingsong Liu, Francesco Iorio, Didier Surdez, Li Chen, Randy J. Milano, Graham R. Bignell, Ah T. Tam, Helen Davies, Jesse A. Stevenson, Syd Barthorpe, Stephen R. Lutz, Fiona Kogera, Karl Lawrence, Anne McLaren-Douglas, Xeni Mitropoulos, "Systematic identification of genomic markers of drug sensitivity in cancer cells ", 483,570–575, 29 March 2012.