ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

A Survey on Cancer Classification of Bio-Molecular Data

M Yasodha, Dr P Ponmuthuramalingam

PhD Research Scholar, Government Arts College (Autonomous), Coimbatore, Tamil Nadu, India

Head, Associate Professor, Government Arts College (Autonomous), Coimbatore, Tamil Nadu, India

ABSTRACT: The microarray technology has enriched the approach of biology research in such a way that scientists can now measure the expression levels of thousands of genes concurrently in a single experiment. Gene expression profiles, which represent the state of a cell at a molecular level, have great medical diagnosis tool. But while compared to the number of genes involved, available training data sets generally have a fairly very small sample size for classification. Feature selection techniques can be used to extract the marker genes which influence the classification accuracy effectively by eliminating the unwanted noisy and redundant genes.

KEYWORDS: Cancer classification, Dimension reduction, Feature selection.

I. INTRODUCTION

Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. Further, it is often the case that finding the correct subset of predictive features is an important problem in its own right. For example, physician may make a decision based on the selected features whether a dangerous surgery is necessary for treatment or not.

Feature selection in supervised learning has been well studied, where the main goal is to find a feature subset that produces higher classification accuracy. Recently, several researches have studied feature selection and clustering together with a single or unified criterion. For feature selection in unsupervised learning, learning algorithms are designed to find natural grouping of the examples in the feature space. Thus feature selection in unsupervised learning aims to find a good subset of features that forms high quality of clusters for a given number of clusters.

Advances in data collection and storage capabilities during the past decades have led to an information overload in most sciences. Researchers working in domains as diverse as engineering, astronomy, biology, remote sensing, economics, and consumer transactions, face larger and larger observations and simulations on a daily basis. Such datasets, in contrast with smaller, more traditional datasets that have been studied extensively in the past, present new challenges in data analysis.

Traditional statistical methods break down partly because of the increase in the number of observations, but mostly because of the increase in the number of variables associated with each observation. The dimension of the data is the number of variables that are measured on each observation.

High-dimensional datasets present many mathematical challenges as well as some opportunities, and are bound to give rise to new theoretical developments. One of the problems with high-dimensional datasets is that, in many cases, not all the measured variables are “important” for understanding the underlying phenomena of interest. While certain computationally expensive novel methods can construct predictive models with high accuracy from high-dimensional data, it is still of interest in many applications to reduce the dimension of the original data prior to any modelling of the data.

Data pre-processing is applied before data mining to improve the quality of the data. Data preprocessing includes data cleaning, data integration, data transformation and data reduction techniques. Cleaning is used to remove noisy data and missing values. Integration is used to extract data from multiple sources and storing as a single repository. Transformation transforms and normalizes the data in a consolidated form suitable for mining.

Reduction reduces the data by adopting various techniques i.e., aggregating the data, attribute subset selection, dimensionality reduction, numerosity reduction and generation of concept hierarchies. The data reduction is also called as feature selection. Feature selection or attribute selection identifies the relevant attributes which are useful to the data mining task.

II. LITERATURE SURVEY

Feature selection is an important pre-processing step in machine learning and data mining. In real-world applications, costs, including money, time and other resources, are required to acquire the features. In some cases, there is a test cost constraint due to limited resources. We shall deliberately select an informative and cheap feature subset for classification. This paper proposes the feature selection with test cost constraint problem for this issue. The new problem has a simple form while described as a Constraint Satisfaction Problem (CSP). Backtracking is a general algorithm for CSP, and it is efficient in solving the new problem on medium-sized data. As the backtracking algorithm is not scalable to large datasets, a heuristic algorithm is also developed. Experimental results show that the heuristic algorithm can find the optimal solution in most cases. We also redefine some existing feature selection problems in rough sets, especially in decision-theoretic rough sets, from the viewpoint of CSP. These new definitions provide insight to some new research directions are performed by Bergenstal*et al* (2014).

Many remote sensing instruments record more channels or bands of data than are actually needed for most applications. As an example, even though the Hyperion sensor on EO-1 produces 220 channels of image data over the wavelength range 0.4–2.4 μm, it is unlikely that channels beyond about 1.0 μm would be relevant for water studies, unless the water were especially turbid. Furthermore, unless the actual reflectance spectrum of the water was essential for the task at hand, it may not even be necessary to use all the contiguous bands recorded in the range 0.4–1.0 μm; instead, a representative subset may be sufficient in most cases are shown by Richards and John (2013).

Bergenstal*et al* (2013) perform the threshold-suspend feature of sensor-augmented insulin pumps is designed to minimize the risk of hypoglycemia by interrupting insulin delivery at a preset sensor glucose value. We evaluated sensor-augmented insulin-pump therapy with and without the threshold-suspend feature in patients with nocturnal hypoglycemia.We randomly assigned patients with type 1 diabetes and documented nocturnal hypoglycemia to receive sensor-augmented insulin-pump therapy with or without the threshold-suspend feature for 3 months. The primary safety outcome was the change in the glycated hemoglobin level. The primary efficacy outcome was the area under the curve (AUC) for nocturnal hypoglycemic events. Two-hour threshold-suspend events were analysed with respect to subsequent sensor glucose values.

Levinskis (2013) describes wavelet transform possible application for convolution neural networks (CNN). As it already known, wavelet transform gives good signal representation in time and frequency domains. This can be useful for CNN input feature reduction as well as architecture simplicity by using only part of coefficients. The result of work is set of experiment which enables to configure out the most appropriate coefficient part. After feature reduction and architecture simplicity achieved configuration could classify data almost ten times faster than original.
Hyper spectral sensors record the reflectance from the Earth's surface over the full range of solar wavelengths with high spectral resolution. The resulting high-dimensional data contain rich information for a wide range of applications. However, for a specific application, not all the measurements are important and useful. The original feature space may not be the most effective space for representing the data. Feature mining, which includes feature generation, feature selection (FS), and feature extraction (FE), is a critical task for hyper spectral data classification. Significant research effort has focused on this issue since hyper spectral data became available in the late 1980s. The feature mining techniques which have been developed include supervised and unsupervised, parametric and nonparametric, linear and nonlinear methods, which all seek to identify the informative subspace. This paper provides an overview of both conventional and advanced feature reduction methods, with details on a few techniques that are commonly used for

analysis of hyper spectral data. A general form that represents several linear and nonlinear FE methods is also presented by Jia*et al* (2013).

Data mining is commonly defined as the extraction of previously unknown and potentially useful information from a database. With the growing volumes of electronic patient records, data mining has become popular to extract hidden patterns in patient data for better understanding of relationships within the data. Data mining in medical domain is unique from that in other domains due to the special characteristics of medical datasets is introduced by Fayyad *et al* (1996).

Paladugu and Sowjanya (2010) introduce medical datasets are often privacy-sensitive, voluminous and heterogeneous with data collected from different sources. The collected data may also need to be characterized mathematically. The rest of the section discusses a few data mining studies that have been conducted in medical and clinical areas. The techniques applied include pattern discovery to identify commonly occurring associations in the dataset, predictive analysis to predict future outcome for a patient based on the existing patient records, and association mining to extract interesting rules from the identified associations. In clinical domain, data mining techniques have been applied to large clinical repositories containing clinical and administrative data collected from electronic sources to identify new disease associations. There have been many recent studies to predict the survival of patients with fatal diseases and to predict treatment outcomes. Studies were conducted by Oztekin et.al. to predict the survivability of heart-lung transplantation patients and by Delen et.al. to predict the survivability of breast cancer patients using prediction models such as neural networks, decision trees , and regression . Decision tree algorithms were also used to effectively predict the survival period of kidney dialysis patients and bladder cancer treatment outcomes. Decision trees based on rules were created and decision making algorithms were used to predict outcomes.

Data mining has also been used for other functions, for instance to fill knowledge gaps in clinical guidelines used in clinical decision support systems. Clinical guidelines hold medical evidence and provide recommendations for clinical conditions that may been countered in practice. Data mining techniques such as decision-tree algorithms were used to extract rules pertaining to the choice of treatment, choice of drugs, etc. from patient records. The rules were extracted for patient subgroups with conditions that were either not addressed in the guidelines or had incomplete rules with missing or imprecise recommended action in the guidelines are shown by Anthony.

Lam et al (2014) performs Breast cancer is traditionally considered as a heterogeneous disease. Molecular profiling of breast cancer by gene expression studies has provided us an important tool to discriminate a number of subtypes. These breast cancer subtypes have been shown to be associated with clinical outcome and treatment response. In order to elucidate the functional consequences of altered gene expressions related to each breast cancer subtype, proteomic technologies can provide further insight by identifying quantitative differences at the protein level. In recent years, proteomic technologies have matured to an extent that they can provide proteome-wide expressions in different clinical materials. This technology can be applied for the identification of proteins or protein profiles to further refine breast cancer subtypes or for discovery of novel protein biomarkers pointing towards metastatic potential or therapy resistance in a specific subtype. In this review, we summarize the current state of knowledge of proteomic research on molecular breast cancer classification and discuss important aspects of the potential usefulness of proteomics for discovery of breast cancer-associated protein biomarkers in the clinic.

Kuruvilla*et al* (2014) performs detection of cancer is the most promising way to enhance a patient's chance for survival. This paper presents a computer aided classification method in computed tomography (CT) images of lungs developed using artificial neural network. The entire lung is segmented from the CT images and the parameters are calculated from the segmented image. The statistical parameters like mean, standard deviation, skewness, kurtosis, fifth central moment and sixth central moment are used for classification. The classification process is done by feed forward and feed forward back propagation neural networks. Compared to feed forward networks the feed forward back propagation network gives better classification. The parameter skewness gives the maximum classification accuracy.

Xiao *et al* (2014) identifying Differentially Expressed (DE) genes from high-throughput gene expression measurements, we would like to take both statistical significance (such as *P*-value) and biological relevance (such as fold change) into consideration. In Gene Set Enrichment Analysis (GSEA), a score that can combine fold change and *P*-value together is needed for better gene ranking.

Non-Gaussian spatial data are very common in many disciplines. For instance, count data are common in disease mapping, and binary data are common in ecology. When fitting spatial regressions for such data, one needs to account for dependence to ensure reliable inference for the regression coefficients. The spatial generalized linear mixed model offers a very popular and flexible approach to modelling such data, but this model suffers from two major shortcomings: variance inflation due to spatial confounding and high dimensional spatial random effects that make fully Bayesian inference for such models computationally challenging. We propose a new parameterization of the spatial generalized linear mixed model that alleviates spatial confounding and speeds computation by greatly reducing the dimension of the spatial random effects. We illustrate the application of our approach to simulated binary, count and Gaussian spatial data sets, and to a large infant mortality data set given by Hughes *et al* (2013).

Learning tasks such as classification and clustering usually perform better and cost less (time and space) on compressed representations than on the original data. Previous works mainly compress data via dimension reduction. In this paper, we propose “double shrinking” to compress image data on both dimensionality and cardinality via building either sparse low-dimensional representations or a sparse projection matrix for dimension reduction by Zhou *wt al* (2013).
Variable and feature selection have turn into the focus of a great deal research in region of request used for datasets with tens or hundreds of thousands of variables are obtainable. These areas include gene expression collection investigation, and combinatorial chemistry. The purpose of changeable selection is three-fold: improving the calculation presentation of the predictors, offering quicker and additional cost-effective predictors, and offering an improved accepting of the fundamental procedure that created the data by Guyon*et al* (2003). Microarray data has been shown by Deutsch (2003) it to be efficacious in individual closely connected cell types that often appear in different forms of cancer, but is not yet realistic clinically. Gene expression profiles may present additional information than morphology and present an alternate to morphology-based tumor categorization systems. Gene selection engage a search for gene subsets that are capable to distinguish tumor tissue from normal tissue, and might have either clear biological understanding or a few inference in the molecular machinery of the tumor genesis. Gene collection is an essential problem in gene expression-based cancer categorization. In the pattern of a discriminate rule, the amount of genes is great relative to the amount of tissue samples. Large genes can harm the presentation of the tumor classification system and enlarge the cost as well. In this report, they talk about criteria and illustrate techniques for dropping the amount of genes and choose a best (or near optimal) subset of genes from an original set of genes for tumor categorization. The realistic advantages of gene selection over additional technique of reducing the dimensionality and amount of genes are given by Xiong*et al* (2001).

III. PROBLEM AND DEFINITIONS

Laboratory instruments become more and more complex and report hundreds or thousands measurements for a single experiment and therefore the statistical methods face challenging tasks when dealing with such high-dimensional data. Feature Selection increases the accuracy of the classifier because it eliminates irrelevant attributes. However, much of the data is highly redundant and can be efficiently brought down to a much smaller number of variables without a significant loss of information. The mathematical procedures making possible this reduction are called dimensionality reduction techniques; they have widely been developed by fields like Statistics or Machine Learning, and are currently a important research topic. In this review, categorize the huge amount of dimension reduction techniques available and give the mathematical insight behind them. Applying feature selection with data mining technique improves the quality of the data by removing irrelevant attributes.

IV. CONCLUSION

This work reviewed, the feature selection techniques that have been employed in cancer classification for the Bio-Molecular data. High dimensionality input and small sample data size are the main two problems that have been triggers the application of feature selection in Bio-molecular data analysis. Numerous and truthful efforts have been conducted during the past several years in the utilization of feature selection to encounter these problems, which mainly can be grouped into three main approaches; filter, wrapper and embedded approaches. The main use of feature reduction is to minimize the huge dimension.

REFERENCES

1. Richards, John A. "Feature reduction." In *Remote Sensing Digital Image Analysis*, pp. 343-380. Springer Berlin Heidelberg, 2013.
2. Bergenstal, Richard M., David C. Klonoff, Satish K. Garg, Bruce W. Bode, Melissa Meredith, Robert H. Slover, Andrew J. Ahmann, John B. Welsh, Scott W. Lee, and Francine R. Kaufman. "Threshold-based insulin-pump interruption for reduction of hypoglycemia." *New England Journal of Medicine* 369, no. 3 (2013): 224-232.
3. Min, Fan, Qinghua Hu, and William Zhu. "Feature selection with test cost constraint." *International Journal of Approximate Reasoning* 55, no. 1 (2014): 167-179.
4. Levinskis, A. "Convolutional Neural Network Feature Reduction using Wavelet Transform." *Electronics and Electrical Engineering* 19, no. 3 (2013): 61-64.
5. Jia, Xiuping, Bor-Chen Kuo, and Melba M. Crawford. "Feature mining for hyperspectral image classification." *Proceedings of the IEEE* 101, no. 3 (2013): 676-697.
6. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." *AI magazine* 17, no. 3 (1996): 37.
7. Paladugu, Sowjanya. "Temporal mining framework for risk reduction and early detection of chronic diseases." PhD diss., University of Missouri--Columbia, 2010.
8. Anthony, Shareen S., and Amit A. Sahu. "Therapeutic Decisions Making System Using Data Mining Techniques: A Review."
9. Lam, S. W., C. R. Jimenez, and E. Boven. "Breast cancer classification by proteomic technologies: Current state of knowledge." *Cancer treatment reviews*40, no. 1 (2014): 129-138.
10. Kuruvilla, Jinsa, and K. Gunavathi. "Lung cancer classification using neural networks for CT images." *Computer methods and programs in biomedicine* 113, no. 1 (2014): 202-209.
11. Xiao, Yufei, Tzu-Hung Hsiao, Uthra Suresh, Hung-I. Harry Chen, Xiaowu Wu, Steven E. Wolf, and Yidong Chen. "A novel significance score for gene selection and ranking." *Bioinformatics* 30, no. 6 (2014): 801-807.
12. Hughes, John, and Murali Haran. "Dimension reduction and alleviation of confounding for spatial generalized linear mixed models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75, no. 1 (2013): 139-159.
13. Zhou, Tianyi, and Dacheng Tao. "Double shrinking sparse dimension reduction."*Image Processing, IEEE Transactions on* 22, no. 1 (2013): 244-257.
14. Guyon, Isabelle, and André Elisseeff (2003), "An introduction to variable and feature selection." The Journal of Machine Learning Research 3: 1157-1182.
15. Deutsch, J. M. "Evolutionary algorithms for finding optimal gene sets in microarray prediction." Bioinformatics 19, no. 1 (2003): 45-52.
16. Xiong, Momiao, Wuju Li, Jinying Zhao, Li Jin, and Eric Boerwinkle (2001), "Feature (gene) selection in gene expression-based tumor classification." Molecular Genetics and Metabolism 73, no. 3: 239-247.