ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2015

A Survey of Advanced Feature Selection Techniques in Genomics

Cephas Paul Edward V[1], Hilda Hepzibah S[2]

Student, Department of Computer Science and Engineering, Anna University, Chennai, India[1]

Asst. Professor, Department of English, AS College, Madurai, India[2]

ABSTRACT: Until today, a numerous feature selection techniques have been put forth and proposed for specific applications in bioinformatics,including genomics and proteomics. For example, consider gene selection. It has been proved to be very useful for biomarker discovery from microarray and mass spectrometry data. Considering the results of the analysis, there exist many supervised feature selection algorithms in genomics whereas only a few unsupervised feature selection algorithms. This paper presents a survey of the recent feature learning techniques available for general purpose and also bio-informatics(genomics) in special.

KEYWORDS: Genomics, Bio-informatics, Feature selection, Proteomics

I. INTRODUCTION

Generally, feature selection, which is also denoted by the terms such as variable selection, attribute selection or variable subset selection. Feature selection is nothing but the process of selecting a subset of relevant features for constructing a suitable model. One of the most critical considerations to be assumed during feature selection process is that the source data may consist of redundant or irrelevant features. Redundant features provide exactly the same information as the features already encountered. Whereas irrelevant features provide no useful information pertaining to any context.

Feature selection is very much varied from feature extraction. Both the terms should not be confused and used in the place of each other. Feature extraction refers to the process of creating new features from functions of the original features. On the other hand, feature selection is the process by which when having a large number of features, chief features of importance are selected and returned. It is like forming a subset from a set and returning the subset. Feature selection techniques are often used in domains where there are many features and only few samples or data points exist. The best example of this is the use of feature selection in analysing DNA microarrays, where there are millions and millions of features, but only a few hundreds of data points or samples. Consider data analysis process. Feature selection also finds its use there. It shows which features are important for prediction and also tries to formulate which features are related with which other and also the nature of their relationship.

In the recent years, feature or gene selection approaches have been more widely utilised in genomics and proteomics inorder to handle a very large amount of data. These data are obtained from techniques such as microarray and mass spectrometry. In the case of microarray studies, a small fraction of the total amount of genes show considerable difference among millions of other genes whose expression levels are also measured simultaneously along with the one under consideration. And so, bassed on the biological phenomenon, it is necessary to characterize the expression profiles of these genes. Gene selection can be useful for multiple situations. First of all, it can considerably save the computational costs involved in subsequent analysis. This, it achieves by reducing the number of genes and thus in turn, improves the prediction performance of classifiers. It makes use of discriminative genes only for the identification of informative genes which are used in the investigation of the biological relevance of those genes.

Gene selection has been proved to be very helpful in the process of biomarker discovery in cancer studies. This involves searching for candidate marker genes which contribute much to the classification of cancer subtypes. This approach is more reliable in the diagnosis point of view. It also paves way for a much better treatment of cancer. Upto now there are a number of techniques that have been put forth for feature selection. And some of these have been

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2015

successfully applied for the multi-featured biological data analysis. Considering the results of the analysis, there exist many supervised feature selection algorithms in genomics whereas only a few unsupervised feature selection algorithms. Unsupervised feature selection is considered greatly useful especially in class discovery. For example, consider clustering. Clustering is usually performed to find clusters or group which are similar and resemble each other in properties in microarray samples. Clustering is performed on this micro-array samples on the on the basis of their expression profiles. But when the results are analysed, the clusters thus obtained are interfered by a huge number of irrelevant genes. Thus one can conclude that such kind of unsupervised feature selection is essential for the exploratory analysis of biological data. Now consider the scenario when class labels are available. When these are provided by external knowledge that might be be unreliable or mislabeled, overfitting can be a potential problem. This overfitting problem can be eradicated by performing feature selection in an unsupervised manner. Usually, it is considered more difficult to identify features that reveal similar sampled groups than that of finding similar patterns across all the samples.

II. GENOMICS

Genomics considered to be a sub-discipline of genetics that applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyse the function and structure of genomes.. It should be also aimed to describe the effect and response to the entire genome's networks.Now in genomics research, after an organism has been selected, three steps are involved. The first one is the sequencing of DNA. This is followed by the assembly of that sequence so obtained which aims at creating a representation of the original chromosome. Finally, the annotation and analysis of that representation of that DNA is involved.

Sequencing

Sequencing in genomes is a similar process like decoding. A genome sequence is considered to be a long sequence of letters in an unknown language. Consider the structural semantics of a natural language. To get the meaning of the sentence, punctuation, capitalization, and every other component of the sentence has its own part or role to contribute to the overall meaning of the sentence. But a genome sequence is like a sentence without any punctuation, capitalization or formatting. It is just a string or sequence of lexemes. Sequencing reveals out hidden interesting data for scientists to work on them. Though, it does not reveal the entire data about the species. It is like a jigsaw puzzle. Scientists generally, break up the sequences and manipulate the bits and pieces obtained. Finally, they build up a new formulation of the sequence. The breaking up is called sequencing.

Assembly

After the genome sequences are broken down and manipulated, they are assembled or built up together forming new sequence in entirely different form. This process is termed as assembly. Multiple fragmented sequence reads must be assembled together on the basis of their overlapping areas.

Annotation

The DNA sequence assembly procedure alone is of only a very little value without any additional analysis involved. Genome annotation is the process of attaching biological information to sequences, and consists of three main steps:

- Identification of portions of the genome that do not code for proteins
- Identification of the elements on the genome with a specialized process called gene prediction
- Attaching biological information to these elements.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2015

III. GENERAL FEATURE SELECTION APPROACHES

Redundancy And Relevancy Analysis Approach

This paper[1] proposes a new and novel technique for feature selection. The proposed approach tries to avoid implicit handling of redundancy in the features. It also incorporates a robust mechanism for the efficient elimination of redundant features. This is achieved using a explicitly handling feature redundancy mechanism. Relevance definitions are put forth and these divide the features into strongly relevant, weakly relevant, and irrelevant classes. The goal of the proposed method in paper is to efficiently find the optimal subset. The authors aim to achieve this goal through a new series of aproaches in feature selection consisting of two steps. First, a relevance analysis is performed which determines the subset of relevant features obtained by removing irrelevant ones. And secondly, redundancy analysis determines and eliminates redundant features from relevant ones and thus produces the final subset. The advantage of the proposed scheme over the other techniques lies in the fact that it decouples both relevance and redundancy analysis. Inability to process image data is its major disadvantage.

Correlation Based Approach

A feature is considered to be good if it is relevant to the class under consideration but is not redundant to any of the other relevant features. If the correlation between two variables is adopted as a measure of goodness of the features, feature selection would be more effective. In other words, if the correlation between a feature and the class is high enough to make it relevant to the class. There exist broadly two approaches to measure the correlation between two random variables. One is based on classical linear correlation and the other is based on information theory.

IV. GENOMIC FEATURE SELECTION APPROACHES

LLDA Based Method

Clearly, it is clearly more challenging to identify features that reveal underlying cluster structures in the samples of genomic data, than to find those exhibiting similar patterns across all the samples. To address this problem, a novel idea has been proposed using an unsupervised feature selection technique. This approach is termed as Laplacian linear discriminant analysis based recursive feature elimination (LLDARFE). It is generally an unsupervised approach for the feature selection process. LLDARFE utilizes a measure called Laplacian score which is also based on graph Laplacian. This measure can be easily applied in an unsupervised manner. The major difference between Laplacian score and LLDARFE is that the former is univariate whereas the latter is multivariate. Thus it is capable of allowing for selecting features that contribute to discrimination in combination with other features.

NN Method for Protein Feature Selection

Artificial Neural Networks(ANN) are a popular tool for classification, prediction and clustering. There are numerous architectures for ANNs existing. ANNs have been found to be useful even in the genomic analysis applications. ANN architecture used is called multilayer perceptron (MLP) with back propagation. The MLP is known to be a powerful function approximator for prediction and classification problems.

Relief Feature Selection

Relief is a novel approach which has proved to be very efficient forestimating feature quality. Gilad-Bachrach et al. have the basis formaulation for Relief. A margin-based criterion measure was employed to assess the quality of the feature sets. The algorithm maintains a data structure known as ‘weight vector’ over all features and updates this vector according to the given sample points(data points). They also explained how to select a determinant threshold in such a way that ensures a low probability that a given irrelevant feature is chosen. The returned weight values allow us to determine which attributes are relevant and to set an order among them. Given a dataset, Relief returns a ranking of features according to an importance weight. This algorithm has an advantage that it has been used in Proteomics providing good results and moreover it is simple and efficient too.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2015

V. CONCLUSION

Thus this paper briefs about genomics and the needs of feature selection in genomics and protein analysis. Several advanced approaches in feature selection for general purposes and feature selection for genomic and bioinformatics purposes have been put surveyed in this paper.

REFERENCES

[1] Niijima, S. and Okuno, Y., Laplacian Linear Discriminant Analysis Approach to Unsupervised Feature Selection, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2009.
[2] Rajdev Tiwari and Manu Pratap Singh, Correlation-based Attribute Selection using Genetic Algorithm, International Journal of Computer Applications, August 2010.
[3] Lei Yu and Huan Liu, Efficient Feature Selection via Analysis of Relevance and Redundancy, Journal of Machine Learning Research, 2004.
[4] Lei Yu and Huan Liu, Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.
[5] Jae-Hong Eom, Byoung-Tak Zhang, Adaptive Neural Network-Based Clustering of Yeast Protein–Protein Interactions, Intelligent Information Technology Lecture Notes in Computer Science Volume 3356, 2005, pp 49-57.
[6] Santi Phithakkitnukoon, Ram Dantu, Inferring Social Groups Using Call Logs, On the Move to Meaningful Internet Systems: OTM 2008 Workshops Lecture Notes in Computer Science Volume 5333, 2008, pp 200-210.
[7] Yongjin Li and Jagdish C. Patra, Selection of Features from Protein-Protein Interaction Network for Identifying Cancer Genes, IEEE 2008.
[8] Jian Zhang , Zongjue Qian , Guochu Shou , Yihong Hu , An Automated On-line Traffic Flow Classification Scheme , Proceedings of the Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2009.