Facioscapulohumeral Muscular Dystrophy Diagnosis Using Hierarchical Clustering Algorithm and K-Nearest Neighbor Based Methodology

Divya Anand, Department of Computer Science and Engineering, Lovely Professional University, Phagwara, India
Babita Pandey, Lovely Professional University, Phagwara, India
Devendra K. Pandey, Lovely Professional University, Phagwara, India

ABSTRACT

The genetic diagnosis of neuromuscular disorder is an active area of research. Microarrays are used to detect the changes in genes for the accurate diagnosis. Unfortunately, the number of genes in gene expression data is very large as compared to number of samples. The number of genes needs to be reduced for correct diagnosis. In the present paper, the authors have made an intelligent integrated model for clustering and diagnosis of neuromuscular diseases. Wilcoxon signed rank test is used to preselect the genes. K-means and hierarchical clustering algorithms with different distance metric are employed to cluster the genes. Three classifiers namely linear discriminant analysis, quadratic discriminant analysis and k-nearest neighbor are used. For the employment of integrated techniques, a balanced facioscapulohumeral muscular dystrophy dataset is taken. A comparative analysis of the above integrated algorithms is presented which demonstrate that the integration of cosine distance metric hierarchical clustering algorithm with k-nearest neighbor has given the best performance measures.

KEYWORDS

Facioscapulohumeral Muscular Dystrophy, Genetic Testing, Hierarchical Clustering, K-Means Clustering, K-Nearest Neighbor, Microarrays

DOI: 10.4018/IJEHMC.2017040103

Copyright © 2017, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
1. INTRODUCTION

Facioscapulohumeral muscular dystrophy (FSHD) is an inherited, autosomal dominant neuromuscular disorder (Tawil, 2008; Lemmers et al., 2012). The term facioscapulohumeral is comprised of the name of muscles of the body i.e. facio (face), scapula (shoulder) and humeral (upper arm). This disease generally affects the upper body and causes the weakness in muscles. It is caused due to the contraction of polymorphic macrosattelite repeat D4Z4 on chromosome 4q35 (Lemmers et al., 2012). According to the FSHD Global Research Foundation, a new estimate says that FSHD affects 1 out of 7500 people and there is no treatment for this disease. The symptoms of the disease are more prone in men than women. For the diagnosis of FSHD, first a neurologist sees the pattern of muscles, which is done using electromyography (EMG). EMG depicts that the person is suffering from muscular dystrophy but not able to differentiate the kind of muscular dystrophy. A physician always prefers the genetic testing for the diagnosis of neuromuscular disorders. This is done by monitoring the expression levels of genes using the microarray technology.

For the genetic diagnosis of FSHD, the blood samples from persons are taken into consideration. In the laboratories, the changes in the chromosomes, DNAs or proteins in those samples are checked. The microarray technology is used to detect these smaller genetic changes. In general, the microarray data of FSHD contains tens of thousands of genes but a very few number of samples. The gene expression matrix is defined using $X = (X_{cd})$. In a gene expression matrix, each row represents a gene and each column represents a sample. The value $X_{cd}$ represents the expression value of a specific gene c in sample d. It is a very complex task to monitor and understand the activity of such a high number of genes to correctly diagnose a disease on time. The informative genes which are related to the disease are very less. As only a few number of genes have different level of activity under condition of interest while most of the genes exhibit a similar expression profile so they are not relevant to the classification task. Hence, to correctly diagnose the disease there is a need to reduce the number of genes. By using the data reduction techniques, we can increase the performance of classifiers by decreasing its computational burden. In addition to that we can reduce the training time, execution time, cost of classification and the risk of overfitting which will help in the easy diagnosis of the disease.

To achieve higher accuracy for the diagnosis of a disease, feature (gene) selection plays a key role in the classification. Feature selection methods are concerned with selecting the small subset of features from the original set of features. These techniques can be applied to both supervised and unsupervised methods. Former, the supervised methods are divided into three main categories: filter methods, wrapper methods and embedded methods. A filter method generates the subset of features while ignoring the communication with the classifiers and works very fast. A wrapper method generates the subsets of features using a specific classification model by taking the feature dependencies into consideration. An embedded method also includes the interaction with the classification model and they are very specific to the learning algorithm.
Challenges and Opportunities of Health Information Exchange (HIE) From Health Consumers’ Perspective
www.igi-global.com/article/challenges-and-opportunities-of-health-information-exchange-hie-from-health-consumers-perspective/204541?camid=4v1a

Theory Driven Organizational Metrics
www.igi-global.com/chapter/theory-driven-organizational-metrics/13079?camid=4v1a