COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 104 (2011) e133–e147

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Application of clustering analyses to the diagnosis of Huntington disease in mice and other diseases with well-defined group boundaries

Jason B. Nikas[a,b,*], *Walter C. Low*[a,c,d]

[a] *Department of Neurosurgery, Medical School, University of Minnesota, Minneapolis, MN, USA*
[b] *Pharmaco-Neuro-Immunology Program, Medical School, University of Minnesota, Minneapolis, MN, USA*
[c] *Graduate Program in Neuroscience, Medical School, University of Minnesota, Minneapolis, MN, USA*
[d] *Department of Integrative Biology and Physiology, Medical School, University of Minnesota, Minneapolis, MN, USA*

ARTICLE INFO

Article history:
Received 3 September 2010
Received in revised form
8 February 2011
Accepted 11 March 2011

Keywords:
Diagnostic methods
Clustering analyses
K-means Clustering
Fuzzy Clustering
Medoid Partitioning Clustering
Hierarchical Clustering
Receiver operating characteristic (ROC) curve analysis
Nuclear magnetic resonance spectroscopy
Metabolomics
Huntington disease

ABSTRACT

Nuclear magnetic resonance (NMR) spectroscopy has emerged as a technology that can provide metabolite information within organ systems *in vivo*. In this study, we introduced a new method of employing a clustering algorithm to develop a diagnostic model that can differentially diagnose a single unknown subject in a disease with well-defined group boundaries. We used three tests to assess the suitability and the accuracy required for diagnostic purposes of the four clustering algorithms we investigated (K-means, Fuzzy, Hierarchical, and Medoid Partitioning). To accomplish this goal, we studied the striatal metabolomic profile of R6/2 Huntington disease (HD) transgenic mice and that of wild type (WT) mice using high field *in vivo* proton NMR spectroscopy (9.4 T). We tested all four clustering algorithms (1) with the original R6/2 HD mice and WT mice, (2) with unknown mice, whose status had been determined via genotyping, and (3) with the ability to separate the original R6/2 mice into the two age subgroups (8 and 12 weeks old). Only our diagnostic models that employed ROC-supervised Fuzzy, unsupervised Fuzzy, and ROC-supervised K-means Clustering passed all three stringent tests with 100% accuracy, indicating that they may be used for diagnostic purposes.

1. Introduction

The ability to utilize common clustering methods in order to develop diagnostic biomarker models that can accurately render a differential diagnosis of a single unknown subject in a given disease state has yet to be demonstrated. Typically, in the vast majority of the cases, common clustering analyses are employed to classify data into two or more groups. For example, in the case of a disease where there are three well-defined groups of subjects (normal, pre-symptomatic, and symptomatic), if a sufficient amount of data of all three

* *Corresponding author at*: Department of Neurosurgery, Medical School, University of Minnesota, 4-218 McGuire Translational Research Facility, 2001 Sixth St., SE, Minneapolis, MN 55455, USA. Tel.: +1 612 625 2868; fax: +1 612 626 9201.
E-mail address: nikas001@umn.edu (J.B. Nikas).
0169-2607/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2011.03.004

of those groups is available, then a common clustering analysis may classify correctly the subjects in the data into three clusters corresponding to the aforementioned three groups. If, on the other hand, a common clustering analysis is presented with the data of a single unknown subject, i.e. it is not known to which of the aforementioned three groups the subject belongs, then, to the best of our knowledge, no common clustering analysis will be able to identify/diagnose that single subject. The ability, therefore, to employ a common clustering analysis in order to develop a diagnostic biomarker model (DBM) that can be used to accurately diagnose a single unknown subject in a disease with well-defined group boundaries constitutes a novel approach. Moreover, and more importantly, this approach has significant implications for the biomedical and clinical sciences because it makes possible the transference of clustering analysis from the research area, i.e. identification of groups in data, to the clinical area, i.e. identification/diagnosis of a single subject.

In two previous studies, using the NMR spectroscopy data we examined mathematical approaches in connection with the identification and assessment of key biomarkers in a disease state, as well as with the development of diagnostic biomarker models and clinical change assessment models [1,2]. In the present study, we investigated four clustering methods (K-means, Fuzzy, Hierarchical, and Medoid Partitioning) that are popular in the medical sciences in connection with the development of diagnostic biomarker models.

Clustering theories first gained popularity in the 1960s, when biologists and social scientists took a keen interest in exploring ways of finding groups in their data [3]. A decade later, aided by advancements in computers, clustering methods were used in medicine, psychiatry, archaeology, anthropology, economics, and finance [4]. Today, there are various clustering methods, K-means, Fuzzy, Hierarchical, Medoid Partitioning, Clustering Regression, etc., and they are used routinely in every scientific field with a focus from the macrocosm to the microcosm—from studying the internal structure of dark matter halos in a set of large cosmological N-body simulations to trying to discover groups of genes in microarray analysis and to predicting protein structural classes [5–12].

As Kaufman and Rousseeuw [3] remarked, "Cluster analysis is the art of finding groups in data." The objective of all clustering methods is to classify N subjects (observations) with P independent variables (IVs) into K clusters according to the spatial relationships among the subjects–subjects in the same clusters are maximally similar, whereas subjects in different clusters are maximally dissimilar. By design, therefore, the primary objective of every clustering method is the correct determination of the number of groups into which a given set of data can be partitioned, and that in itself constitutes, by far, the most difficult task that a clustering analysis has to do. As it so happens, in the medical sciences, especially in the area of diagnostics, the number of groups is known. In our study of experimental Huntington disease (HD), for example, we know a priori that we have two, and only two, groups of mice: a mouse can be either a normal wild type (WT) or an R6/2 (HD) mouse. We can therefore ask a clustering method not to waste time examining all the possible clustering outcomes but to focus instead in classifying all of our subjects (data) into only two clusters. This constitutes a significant bypass of the most difficult course – in terms of both obstacles and potential pitfalls – that a clustering analysis has to traverse. This holds true for any other disease where the number of groups is known. If, for example, we studied a disease with three groups, including the normal group, then we would pre-set the number of clusters to three.

Availing ourselves of the aforementioned significant theoretical advantage, we sought to answer the question of whether it was possible to use a common clustering method to ultimately render a differential diagnosis of a single unknown subject in a disease with well-defined group boundaries. To address this question, we first developed a clustering approach that made it possible to use a common clustering method for such a purpose, and we subsequently investigated four clustering methods (K-means, Fuzzy, Hierarchical, and Medoid Partitioning) by applying them to the *in vivo* analysis of the striatal metabolomic profile of R6/2 transgenic mice with Huntington disease (HD) and WT mice using proton nuclear magnetic resonance (^{1}H NMR) spectroscopy. We first assessed the clustering models in an unsupervised way. Then, we introduced the concept of employing ROC curve analysis with the express purpose of supervising the clustering models in order to increase their accuracy, and we subsequently assessed the performance of the ROC-supervised clustering models and compared it to that of their unsupervised counterparts. Our ultimate goal was to accomplish the following two objectives:

(1) Construct diagnostic biomarker models (DBMs) that could accurately diagnose R6/2 mice as a prototype for the diagnosis of diseases. Since HD is a neurodegenerative disease with well-defined group boundaries (WT vs. R6/2), since genotyping is available for the R6/2 mice, and since genotyping is the gold standard, we would use HD to test the accuracy of our DBMs.
(2) Subject all four clustering methods to thorough, stringent tests in order to assess their strengths, weaknesses, and overall suitability for diagnostic applications.

2. Methods

2.1. *Animal methods*

2.1.1. *R6/2 transgenic mice*

Original R6/2 mice were purchased from the Jackson Laboratories (Bar Harbor, ME, USA) and bred by crossing transgenic males and wild type (WT) females at 5 weeks of age. Offspring were genotyped according to established procedures [13] and the Jackson Laboratory. All animal breeding and all animal experiments described in this study were performed in accordance with the procedures approved by the University of Minnesota Institutional Animal Care and Use Committee. This study was specifically approved by the aforementioned Institutional Committee.

2.1.2. *Animal preparation*

Prior to the *in vivo* ^{1}H NMR (proton nuclear magnetic resonance) scanning, all animals were anesthetized and