Classification Methods for HIV-1 Medicated Neuronal Damage

Mengjun Wang[1], Jialin Zheng[2], Zhengxin Chen[3*], and Yong Shi[3,4]
[1]*Heart & Vascular Institute, Henry Ford Health System*
2799 West Grand Boulevard, Detroit, MI 48202
[2]*Center for Neurovirology and Neurodegenerative Disorders*
University of Nebraska Medical Center, Omaha, NE 68198-6880
[3]*College of Information Science and Technology*
University of Nebraska at Omaha, Omaha, NE 68182
[4]*Chinese Academy of Sciences Research Center on Data Technology & Knowledge Economy,*
Graduate University of the Chinese Academy of Sciences,
Beijing 100080, China
* zchen@mail.unomaha.edu (Corresponding author)

Abstract

HIV-1-associated dementia (HAD) is the most devastating disease happened in the central nervous system of AIDS patients. Neuronal damage, the early indicator of HAD, under different treatments can be applied to design and study specific therapies for the prevention or reversal of the neuronal death associated with HAD. A computer-based image program was used to quantitatively estimate the change of neurites, arbors, branch nodes, and cell bodies in cultured cortical neurons. Nine attributes (variables) and two classes G2 (non-treatment control group) and G4 (gp120-treatment group) were considered to describe the statuses of neuronal damage. Various classification methods have been carried out in our research group. In this paper, we focus on using logistic regression method for classification, and compare the resulting predictive accuracy with that of using two-class multiple criteria linear programming (MCLP) and neural networks (NN) models conducted earlier. The results show that logistic regression obtained the best classification accuracy. As a pilot study, it demonstrates the use and effectiveness of statistical method in the classification mining of neuronal damage associated with HAD.

1. Introduction

HIV-1-associated dementia (HAD) is the most devastating disease happened in central nervous system of AIDS patients. In the last several years, various classification methods have been carried out in our research group led by one of the authors (JZ). In this paper, we focus on using logistic regression method for classification, and compare the resulting predictive accuracy with that of using two-class multiple criteria linear programming (MCLP) and neural networks (NN) models conducted earlier. Below we provide necessary background of our research, a brief description of the method used in this research, as well as a brief summary of the result.

2. Background

The histopathological effect related to HAD is HIV-1 encephalitis (HIVE). Damage of dendritic arbor and reduction of synaptic density provide early change images of the neuronal cell. The pathogenesis of HAD is not clear by now [1]. Brain mononuclear phagocytes (MP) are activated by HIV-1 and are believed to produce main mediators of neuronal injury in HAD. Normally MP has the protective function to human body, but MP produces neurotoxic factors based on the different stages. HIV-1 gp120, a viral neurotoxic protein, is also secreted by infected brain MP. Note that gp120 could interact with receptors on neurons and interfere with cell signaling leading to neuronal damage or it could also indirectly induce neuronal injury through the production of other neurotoxins. However, it is still unclear how MP evolves from producing neurotrophins to producing neurotoxins [1].

Neural cell damage could be an early indicator of HAD. The early stages of damages to neurons have possibility to be reversed. The ability to predict neuronal death through identification of these early damages will be valuable in fighting with this debilitating disorder. Also the ability to identify neuronal damage resulting from HAD is crucial for designing specific therapies for the treatment of HAD.

To predict neuronal loss, a quantitative measurement of neuronal injury in neuronal process was developed by one of the authors' lab (JZ) to discover neuronal morphology changes quantitatively in the neurotic network of cortical neurons [1]. The findings of neurons

in the experiment are believed as an early predictor of HAD. Based on a database obtained by *ImagePro Plus* software on the fixed neuronal cultures, nine major variables of neuronal damage and dropout were selected in HIV-1-associated dementia from the observations. These include (i) *Neuri*, the number of neurites, (ii) *Arbor*, the number of arbors, (iii) *BraNod*, the number of branch nodes, (iv) *Avglen*, the average length of arbors, (v) *NeArRat*, the ratio between the number of neurites and the number of arbors, (vi) *Area*, the area of cell bodies, (vii) *MaxLen*, the maximum length of the arbors, (viii) *CulT*, the cell culture time, and (ix) *TreT*, the cell treatment time. Neurites are the single lines of a neuron except the single axon. Arbors are the groups of neurites extending from a single point on the cell body. An arbor containing more than one neurite means that it is branched. Branch nods are the points at which neurite segments branch into separate neurites. The area of a cell body represents the size of a cell body. The maximum length of the arbors is the longest length of the arbor. Average arbor lengths and ratios between the number of neurites and the number of arbors were also calculated. The cell culture time is the duration when the neuron grows normally and extrinsic factor has not been added to affect growth, and the cell treatment time is the duration when the neuron was growing under the effects of extrinsic factor.

3. Methods used for conducting the research

Originally four classes were considered to describe the statuses of evolution from neurotrophins to neurotoxins. Because the analysis of {non-treatment group (G2) vs. gp120 treatment group (G4)} is the most valuable identification from the biological perspective [1], we have focused our research on the classification of this pair. The knowledge discovered in neuronal damages under the classes can be applied to design and study specific therapies for the prevention or reversal of the neuronal demise associated with HAD [1].

Numerous methods have been developed for classification task for intelligent data analysis and data mining, each has its own strength and weakness. In the work related to this pilot study, our bioinformatics research group has applied two data mining methods [1], a two-class model of multiple criteria linear programming (MCLP) and neural network (NN). The acquired results showed that two classification methods predicted and uncovered the data patterns among four class pairs, and MCLP provided a more comprehensive description of changes of HIV-1 medicated neuronal damage. But an obvious identification of class pair {G2 vs. G4} was not obtained. Especially in the NN model, 42%, the value of *Specificity* (the correct identification rate of G4) in test group can be regarded as a failure in terms of the acceptable rate 55%. We applied this traditional statistical method to develop new classification model and applied its advantages to understand the complex data structure.

4. Summary of results

The study was conducted in the following manner. (1) The logistic regression classification model was developed and compared with other two models (MCLP and NN) to test for the potential of the series data mining application in HAD research. The predictive accuracy was estimated according to principles of data mining. The logistic regression model showed exciting advantages in both predictive accuracy and computing time. It indicates that logistic regression is a competitive classification method in the neuronal injury research. (2) Logistic regression was further used to determine the contribution of each variable to the prediction of group. The model clearly indicates the size and direction of *odds* change based on 1-unit change of variable. The *odds* of an event is the ratio of the expected numbers of times that an event will occur to the expected number of times it will not occur. For the final effect, the number range of variables should be taken into account. Based on 1-unit change, we found that *NeArRat* has the greatest contribution on the prediction of G2 and G4. *Neuri*, *Arbor*, *CulT* and *AvgLen* have the positive contribution on the prediction of G2. However, *TreT* has a negative effect on the prediction of G2. These results are consistent to the actual biological meanings. We can delete less important attributes in our future study, including *BraNod* (which has been discovered as a linear combination of *Neuri* and *Arbor*) and *Area* (because it does not play a role in the class identification). The results also indicate *MaxLen* has a little negative effect on the prediction of G2, which is reverse to the effect of *AvgLen*. This can not be explained from the biological perspective. It suggests that *MaxLen* is not a reliable measurement of arbor length. In contrast, *AvgLen* is a more variable and reliable measurement. The logistic regression builds the practical basis for finding the potentially important variables from the complex structure of data. The results can conduct our future studies.

References

[1] J. Zheng, W. Zhuang, N. Yan, G. Kou, H. Peng, C. McNally, D. Erichsen, A. Cheloha, S. Herek, C. Shi and Y. Shi, Classifications of HIV-1 Medicated Neuronal Dendritic and Synaptic Damage using Multiple Criteria Linear Programming, *Neuroinformatics*, 2004.