CSIAM Trans. Life Sci.
doi: 10.4208/csiam-ls.SO-2024-0004a

Vol. **1**, No. 1, pp. 134-152
March 2025

Disease Prediction with a Maximum Entropy Method

Michael Shub[1], Qing Xu[2,*] and Xiaohua Xuan[2]

[1] *Math Department, City College and the Graduate Center of CUNY, New York 10031, USA.*
[2] *UniDT, Shanghai 200436, China.*

Received 18 June 2024; Accepted 30 October 2024

Abstract. In this paper, we propose a maximum entropy method for predicting disease risks. It is based on a patient's medical history with diseases coded in International Classification of Diseases, tenth revision, which can be used in various cases. The complete algorithm with strict mathematical derivation is given. We also present experimental results on a medical dataset, demonstrating that our method performs well in predicting future disease risks and achieves an accuracy rate twice that of the traditional method. We also perform a comorbidity analysis to reveal the intrinsic relation of diseases.

AMS subject classifications: 00A69, 92B10

Key words: Disease prediction, maximum entropy, bioinformatics.

1 Introduction

Disease prediction is an effective way to assess a person's health status. Studies [1,3] have shown that in many cases, there are identifiable indicators or preventable risk factors before the onset of the patient's disease. These early warnings can effectively reduce the individual's risk of disease. Theoretically, this can reduce the number of treatments needed and increase the necessary effective interventions. However, the combination of problem factors caused by different diseases and the patient's past medical history are so complicated that no doctor can fully understand all of this. Currently, doctors can use family and health history and physical examinations to estimate the patient's risk and guide laboratory tests to further evaluate the patient's health. However, these sporadic and qualitative "risk assessments" are usually only for a few diseases, depending on the experience, memory and time of the particular doctor. Therefore, the current medical care is after the fact. Once the symptoms of the disease appear, it is involved, rather than actively treating or eliminating the disease as soon as possible.

*Corresponding author. *Email addresses:* shub.michael@gmail.com (M. Shub), qing.xu@unidt.com (Q. Xu), michael.xuan@unidt.com (X. Xuan)

Today the prevailing model of prospective heath care is firmly based on the genome revolution [14,16]. Indeed, technologies ranging from linkage equilibrium and candidate gene association studies to genome wide associations have provided an extensive list of disease-gene associations, offering us detailed information on mutations, single nucleotide polymorphisms (SNPs), and the associated likelihood of developing specific disease phenotypes [10,18]. The basic assumption behind the research is that once we have classified all disease-related mutations, we can use various molecular biomarkers to predict each individual's susceptibility to future diseases, thus bringing us into a predictive medicine era [2]. However, these rapid advances have also revealed the limitations of genome-based methods. Considering that the signals provided by most disease-related SNPs or mutations are very weak, it is becoming increasingly clear that the prospect of genome-based methods may not be realized soon [4,9]. Does this mean that prospective disease prediction methods must wait until genomics methods are sufficiently mature? Our purpose is to prove that the method based on medical history provides hope for the prospective prediction of disease.

In this paper, we mainly study the disease prediction and comorbidity of diseases. Our approach is distinctly different in that we are trying to build a general predictive system which can utilize a less constrained feature space by taking into account all available demographics and previous medical history. Moreover, we rely primarily on International Classification of Diseases, tenth revision, Clinical Modification (ICD-10) codes (see Section 2) for making predictions to account for the previous medical history, rather than specialized test results.

2 Data

Our database comprises the medical records of 354,552 patients in China with a total of 2,904,257 hospital visits. The data was originally compiled from Insurance claims during 2007 to 2017. Such medical records are highly complete and accurate, and they are frequently used for epidemiological and demographic research.

The input for our methods consists of each patient's personal information, such as gender, birthday, treatment-date, and diagnosis history, provided per patient's visit. Each data record consists of a hospital visit, represented by a patient ID and a diagnosis code per visit, as defined by the International Classification of Diseases, tenth revision, Clinical Modification. The International Statistical Classification of Diseases and Related Health Problems provides codes to classify diseases and a wide variety of signs, symptoms, abnormal findings, social circumstances, and external causes of injury or disease. It is published by the World Health Organization. Each disease or health condition is given a unique code, and can be up to 6 characters long, such as A01.001. The first character is a letter while the others are digits. ICD-10 codes are hierarchical in nature, so the 6 characters codes can be collapsed to fewer characters identifying a small family of related medical conditions. For instance, code A01.001 is a specific code for typhoid fever. This code can be collapsed to A01.