GENETIC ALGORITHM IMPROVEMENTS TO FEATURE OPTIMIZATION FOR CLASSIFICATION OF MEDICAL RECORDS

Yuda Syahidin, Ari Sakti Prasetya
Manajemen Informasi Kesehatan, Piksi Ganesha Politechnic, Bandung, Indonesia

Abstract

Feature selection in medical records is necessary because the data usually contains many irrelevant features and noise. Electronic Health Record, abbreviated as EHR, makes it possible to analyze large amounts of medical data. A Genetic Algorithm is widely used for feature selection because it has the ability or potential for global optimization of the selected features. Genetic Algorithm-based methods include many iterations (generations) in the crossover process, and mutation can produce new individuals because the Genetic Algorithm adopts a fitness value to represent how "good" the resulting individual is. The problem with heuristic algorithms is those simple genetic algorithms are not for processing high- dimensional data. Genetic algorithms in solution search techniques always get local optimum solutions which can cause failure to obtain optimal solutions during random searches. From these limitations, developing and improving genetic algorithms for feature selection on clinical data is necessary. First, sort the features based on the feature evaluation criteria to exclude irrelevant features through the fitness process in the evaluation with the accuracy value of the Support Vector Machine calculation. This way reduces the number of features and results in optimal features. Then to get the optimal solution, it is necessary to optimize the subset features that have been selected using a machine learning algorithm that determines the best parameters using a genetic algorithm.

Keywords: Electronic Medical Record, Genetic Algorithm, Feature Optimization, Support Vector Machine

Introduction

Genetic Algorithms are widely used for feature selection because they have the ability or potential for global optimization of the selected features (Dewan & Sharma, 2015; Ghorbani dkk, 2020; Kalinin dkk, 2018; Pawlovsky & Matsuhashi, 2017).

Genetic Algorithm-based methods include many iterations (generations) in the process of crossover, and mutation can produce new individuals because the Genetic Algorithm adopts a fitness value to represent how "good" the resulting individual (Duan dkk, 2021; Guo dkk, 2014). Each individual is evaluated through the stages of fitness. Calculation of fitness can be done with the MAE (Mean Absolute Error) technique stages (Guan dkk, 2017; Nguyen dkk, 2021), which is the output of the prediction model with a combination of specification features that act as model inputs (Dewan & Sharma, 2015; Liang dkk, 2020).

Research using genetic algorithms, the search technique carried out, namely sequence searches, can be divided into

three categories such as forward search, reverse search and two-way search (Dewan & Sharma, 2015; Ghorbani dkk, 2020; Majidnezhad, 2015; Nguyen dkk, 2021). Time complexity is the optimization chosen based on the lowest, but the resulting subset of features only produces the locally optimal (Sun dkk, 2004), Random search can make the advantages of two methods; namely, The temporal complexity is lower than the global ideal but greater than the sequence search, and the best solution can be calculated. The genetic algorithm is a popular random search technique that has been and continues to be used in a variety of fields. (Guo dkk, 2014; Liang dkk, 2020; Luque dkk, 2011) presents a genetically based algorithm to perform simultaneous parameter selection and optimization for features from clinical data. The ideal global search, sequence search, and random search are the three types of search feature subsets and search algorithms. Global optimum tracking identifies a global subset of the original feature collection that is optimal. So far, only the whole method and the branch-and-bound (BBM) method have been implemented.(Peng dkk, 2021), It can attain global optimality, but is only useful in low-dimensional areas because to its high time complexity (Daniel, 2018; Zhaoke Huang dkk, 2019).