2 col example

Introduction

Biomarker plays an important role in early disease diagnosis including cancer. The World Health Organization defines a biomarker as any structure or process in the body that is measurable and affects the prognosis or outcome of the disease. Today, biomarkers can be identified using bioinformatics tools. The detection of biomarkers in the field of bioinformatics is considered more as a problem of feature selection. Many feature selection algorithms have been used for biomarker discovery however these algorithms do not have enough accuracy or have computational complexity. For this reason, the researchers discard the high accuracy algorithms because they are time consuming. We redesigned an efficient algorithm based on parallel algorithms. We used the Cancer Genome Atlas (TCGA) including breast cancer patients. The proposed algorithm has the same accuracy and increases the speed of

Risk Factors

- Prolonged sun exposure
- Family history of melanoma
- Immunosuppression
- Exposure to tanning beds

Diagnostic Steps

1. Visual inspection under natural light
2. Dermatoscopic evaluation
3. Photographing for follow-up

4. Biopsy if suspicious characteristics are observed

Clinical Recommendation

Any matter, structure, or process that is measurable in the body and can affect the prediction or trend of the disease is known as a biomarker [1, 2]. Biomarkers can diagnose the disease before clinical symptoms appear, so it is vital to use them in early detection of diseases [3]. The discovery of cancer biomarkers results in the early detection of cancer, which will have a significant impact on the mortality rate of the disease. Biomarkers are obtained from the analysis of biomolecules such as DNA, RNA and proteins and can themselves be proteins, genes, hormones and enzymes [4]. The discovery of biomarkers is a matter of feature selection in bioinformatics, especially when the distinction between features is important [5]. Feature selection means finding a subset of attributes with the minimum possible size that contain the necessary information for the intended purpose. In the biomarker detection problem, we also encounter a large number of features and samples. The goal is to select a subset of the minimal features that are very close and efficient to the examples. The feature selection algorithms, in addition to returning a set of features as output, reduce the dimension and the redundancy of data and increase the accuracy [6]. With the help of these algorithms, it is possible to identify and validate biomarkers in several steps. First, the needed genomic and proteomic data is collected and organized by databases. In addition, unnecessary data is deleted. The second step involves feature selection and, in some cases, use classification methods. Appropriate tools should validate the candidate biomarkers at a later stage [7].

Feature selection algorithms can be grouped into three categories: filter, wrapper, and embedded. Filter approach work based on the inherent nature of the data. These types of algorithms are usually simple and do not have high computational complexity. However, these algorithms are not very accurate and usually not stable. It means, for each execution, it usually returns different attributes [8]. An example of these algorithms is the Relief algorithm that is unstable and has been used for the detection of biomarkers [9]. Wrapper algorithms use classification methods for ranking features. Wrapper algorithms have high computational complexity and are not sufficiently fast, but due to using classification, the relationships between features are considered and algorithms are highly accurate [8]. One of the most widely used algorithms for biomarker discovery is support vector machine that was used as a wrapper algorithm [10]. Because this algorithm has constraints as a feature selection algorithm, a combination of this algorithm and a recursive feature elimination algorithm (SVM-RFE1) was proposed. This algorithm has increased the accuracy of its previous algorithms [11]. Since then, the SVM-RFE was considered as a benchmark algorithm for other feature selection classification [12]. Feature selection algorithms generally have different performances on different data types in terms of biology [8]. In addition, due to the growing amount of data in biology area, the choice of fast algorithms has priority. Only the filter algorithms with this data volume can return the output at an acceptable time. As stated, these algorithms do not have the required precision in detecting cancer biomarkers, while misinterpreting biomarkers can cause many problems [8]. The use of wrapper algorithms is now increasing rapidly due to the high accuracy but the computational complexity of these algorithms and their running time is significant [12]. One of the main solutions to reduce the computational complexity of algorithms is the parallelization method. Parallelization means increasing the number of processors and dividing the work or data into several sub-tasks so, the calculations are divided between the processors and the response time is reduced [13].

The SVM-RFE algorithm is one of the best feature selection algorithms for biomarker detection and is considered by many researchers in this field. This algorithm is stable, meaning that the output of the algorithm is constant at each repeat. It also has good accuracy in detecting biomarkers, and it is very time-consuming just because of the use of a support vector machine algorithm [14, 15]. This algorithm performs very well on genomic data [12]. In this paper, first, the dataset is introduced then the parallelization method for this algorithm is described.

The Cancer Genome Atlas (TCGA) is a large dataset of genomic variation in more than 33 types of cancer, which is valuable for computing tools. This dataset contains the genomic changes, DNA sequences, and gene expression in a cell tumor relative to a healthy cell. This study covers gene expression for more than a thousand patients with breast cancer and includes the expression of genes, metabolism, and clinical data for 2012-2015 [14]. By integrating these

In 2005, a different implementation of the SVM-RFE algorithm was introduced which at each step instead of one support vector machine, it uses several support vector machines [15]. This algorithm was known as a tool to select the effective genes in cancer classification and it returned better features than the SVM-RFE algorithm. In addition, the quality of the classifier is better [15]. Due to using several classes at the classification stage, this algorithm was called the mSVM-RFE algorithm. The algorithm receives the data from the user, then Fold method is used to create sub- samples randomly. The user with k variable introduces the number of these subsamples. In the second step, the support vector machines (the number is specified by the user) are trained on each subsample, and the value of each feature is estimated using all machines. Each machine has its weight vector and for each property such as let be the weight of the feature SVMs [15]: And the ranking score for each feature such as , is defined as: In the next step, the feature with the smallest ranking is eliminated.

The second and third steps are repeated until for all features, their ranking be calculated. In the end, five (or more) features with the highest ranking are returned as an output [15]. In this algorithm, as in SVM-RFE, in each iteration, we can eliminate several features instead of one by one. This algorithm is more expensive than the SVM-RFE algorithm because it uses multiple SVMs instead of one machine. However, this cost will ultimately lead to a better feature selection and more accurate ranking. In addition, one of the ways that we can increase the stability of algorithms is to select the features on different subsets [15]. Both algorithms were performed on the breast, lung, colon, and blood dataset and the best features for breast cancer and blood cancer were obtained by mSVM-RFE. On colon cancer, the selected features by SVM-RFE were better [15].