Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

An unsupervised machine learning method for discovering patient clusters based on genetic signatures

Christian Lopez[a], Scott Tucker[b,c], Tarik Salameh[b], Conrad Tucker[a,d,e,*]

[a] Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802, USA
[b] Hershey College of Medicine, The Pennsylvania State University, Hershey, PA 17033, USA
[c] Engineering Science and Mechanics, The Pennsylvania State University, University Park, PA 16802, USA
[d] Engineering Design Technology and Professional Programs, The Pennsylvania State University, University Park, PA 16802, USA
[e] Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA

ARTICLE INFO

Keywords:
Unsupervised machine learning
Clustering analysis
Genomic similarity
Multiple sclerosis

ABSTRACT

Introduction: Many chronic disorders have genomic etiology, disease progression, clinical presentation, and response to treatment that vary on a patient-to-patient basis. Such variability creates a need to identify characteristics within patient populations that have clinically relevant predictive value in order to advance personalized medicine. Unsupervised machine learning methods are suitable to address this type of problem, in which no *a priori* class label information is available to guide this search. However, it is challenging for existing methods to identify cluster memberships that are not just a result of natural sampling variation. Moreover, most of the current methods require researchers to provide specific input parameters *a priori*.

Method: This work presents an unsupervised machine learning method to cluster patients based on their genomic makeup without providing input parameters *a priori*. The method implements internal validity metrics to algorithmically identify the number of clusters, as well as statistical analyses to test for the significance of the results. Furthermore, the method takes advantage of the high degree of linkage disequilibrium between single nucleotide polymorphisms. Finally, a gene pathway analysis is performed to identify potential relationships between the clusters in the context of known biological knowledge.

Datasets and results: The method is tested with a cluster validation and a genomic dataset previously used in the literature. Benchmark results indicate that the proposed method provides the greatest performance out of the methods tested. Furthermore, the method is implemented on a sample genome-wide study dataset of 191 multiple sclerosis patients. The results indicate that the method was able to identify genetically distinct patient clusters without the need to select parameters *a priori*. Additionally, variants identified as significantly different between clusters are shown to be enriched for protein-protein interactions, especially in immune processes and cell adhesion pathways, via Gene Ontology term analysis.

Conclusion: Once links are drawn between clusters and clinically relevant outcomes, Immunochip data can be used to classify high-risk and newly diagnosed chronic disease patients into known clusters for predictive value. Further investigation can extend beyond pathway analysis to evaluate these clusters for clinical significance of genetically related characteristics such as age of onset, disease course, heritability, and response to treatment.

1. Introduction

With advancements in genome-wide association study (GWAS) techniques and the advent of low cost genotyping arrays, researchers have developed a significant interest in applying Machine Learning (ML) methods to mine knowledge from patients' genomic makeup [1,2]. This knowledge has allowed researchers to improve gene annotation and discover relationships between genes and certain biological phenomena [3,4].

The fields of personalized and stratified medicine benefit greatly from ML. For example, many cases in the field of pharmacogenetics have identified genetic variants with clinically actionable impacts on drug response and metabolism [5,6]. Moreover, many chronic disorders (e.g., asthma, diabetes, Crohn's disease) have genomic etiology, clinical presentation, and response to treatment that vary on a patient-to-patient basis. Such variability reveals a need to identify characteristics

[*] Corresponding author at: 213 N Hammond Building, University Park, PA 16802, USA.
E-mail address: ctucker4@psu.edu (C. Tucker).

https://doi.org/10.1016/j.jbi.2018.07.004
Received 21 December 2017; Received in revised form 22 June 2018; Accepted 7 July 2018

within patient populations that have clinically relevant insights. For example, Multiple Sclerosis (MS) is a chronic inflammatory disorder in which progressive autoimmune demyelination and neuron loss occur in the central nervous system. MS varies from patient-to-patient in genomic etiology, disease progression, clinical presentation, and response to treatment. Hence, MS patients, like other chronic autoimmune patients, could benefit from ML methods that advance personalized medicine.

Machine learning methods are commonly classified into *supervised* and *unsupervised* methods. Supervised methods, such as Support Vector Machines [7] and Random Forests [8,9], have been extensively used in the field of bioinformatics. These methods classify new objects to a determinate set of discrete class labels while minimizing an empirical loss function (e.g., mean square error). However, supervised methods require the use of a training set that contains *a priori* information of several objects' class labels. In contrast, unsupervised methods do not require a training set that contains *a priori* information of objects' class labels as input. Unsupervised methods are able to detect potentially interesting and new cluster structures in a dataset. Moreover, they can be implemented when class label data is unavailable. Hence, if the objective of a study is to discover the class labels that best describe a set of data, unsupervised machine learning should be implemented in place of supervised methods [2]. However, it is challenging for existing unsupervised ML methods to identify object memberships that are due to the underlying cluster structures in the dataset, rather than the results of natural sampling variation [10]. Moreover, most current methods require researchers to provide certain input parameters *a priori* (e.g., number of clusters in the dataset), which can limit their applicability.

In light of the limitations of existing methods and the need to advance personalized medicine, an unsupervised machine learning method to cluster patients based on their genomic similarity is presented. The method integrates statistical analysis that accounts for *family-wise-error* rate, allowing the method to identify clusters resulting from the underlying structure of the data and not just due to random chance. Moreover, the method takes advantage of the high degree of linkage disequilibrium between Single Nucleotide Polymorphisms (SNP) by pruning correlated nearby SNPs, which helps reduce redundant variants in the dataset. Finally, a gene pathway analysis shows the potential relationships between the clusters in the context of known biological knowledge. The proposed method is capable of clustering patients based on their genomic similarity without *a priori* information. Moreover, it is capable of identifying the significant variants (i.e., SNPs) between patient sub-groups within a cohort with a common disorder. Successfully identifying distinct genetic subtypes of patients within genomic datasets demonstrates the potential of this method to advance personalized medicine of complex diseases with heritable components, especially autoimmune disorders which have many shared susceptibility loci [11].

2. Literature review

In the last decade, the field of bioinformatics has seen a significant number of publications implementing unsupervised machine learning methods, such as clustering algorithms [12–14]. Clustering algorithms partition data objects (e.g., genes, patients) into groups (i.e., clusters), with the objective of exploring the underlying structure on a dataset [15]. In the medical field, these algorithms have been implemented to identify sets of co-expressed genes [16], compare patients' prognostic performance [17], cluster patients based on their medical records [18], and identify subgroups of patients based on their symptoms and other variables [19].

In previous work, genomic stratification of patients (i.e., stratified medicine) has been able to match specific therapy recommendations to genetic subpopulations by predicting therapeutic response [5,6]. However, most of these studies implemented class label data (i.e., response to treatment) to cluster patients. In clinical datasets, class label information is not widely available for convenient patient clustering. Unsupervised machine learning methods can be used in such cases to identify clusters within the dataset. Further investigation of genetic subgroups within a cohort of patients can offer a better clinical prediction of age of onset, disease course, heritability, and response to therapy, leading to improved outcomes [20].

2.1. Hierarchical clustering algorithms

Agglomerative hierarchical clustering algorithms are one of the most frequently used algorithms in the biomedical field [21,22]. Researchers have found that hierarchical clustering algorithms tend to perform better than other algorithms (e.g., k-means, partitioning around Medoids, Markov clustering) when tested on multiple biomedical datasets [23]. The objective of any agglomerative hierarchical clustering algorithm is to cluster a set of n objects (e.g., patients, genes) based on an $n \times n$ similarity matrix. These clustering algorithms have grown in popularity due to their capability to simultaneously discover several layers of clustering structure, and visualize these layers via tree diagrams (i.e., dendrogram) [10]. Even though these algorithms allow for easy visualization, they still require preselecting a similarity height cut-off value in order to identify the final number of clusters. In other words, it still requires researchers to know *a priori* the number of clusters in the dataset.

Agglomerative hierarchical clustering algorithms can be implemented with different linkage methods. For example, Ahmad et al. [17] implemented the Ward's linkage method to compare patients' prognostics performance; while Hamid et al. [19] implemented the Complete linkage method to identify unknown sub-group of patients. Unfortunately, depending on the underlying structure of the data, different clustering results can be obtained by implementing different linkage methods. Ultsch and Lötsch [24] demonstrated that neither the Single nor Ward's linkage methods provided similar clustering results when tested with the Fundamental Clustering Problem Suite (FCPS) datasets [25]. Their results reveal that these linkage methods were able to correctly cluster all the objects in only a subset of the FCPS datasets. Similarly, Clifford et al. [26] discovered that while testing multiple simulated GWAS datasets, the linkage methods of Median and Centroid were the only ones to consistently be outperformed by the Single, Complete, Average, Ward's, and McQuitty methods. In light of these, Ultsch and Lötsch [24] proposed the use of emergent self-organizing maps to visualize clustering of high-dimensional biomedical data into two-dimensional space. Even though, their method allowed for better visualization, it still required preselecting the number of clusters as well as other parameters to perform correctly (e.g., toroid grid size) [24].

2.2. Parameter selection in clustering algorithms

In order to avoid preselecting input parameters *a priori* (e.g., the number of clusters), researchers have implemented cluster validation metrics. For example, Clifford *et al.* (2011) [26] proposed a method that aimed to capture the clustering outcome of multiple combinations of linkage method and similarity metric based on the Silhouette index [27]. The Silhouette index was used to rank the results of the clustering combinations, and select the best cluster set (i.e., cluster set with largest average Silhouette index). Similarly, Pagnuco et al. [16] presented a method that implemented several linkage methods and implemented modified versions of the Silhouette and Dunn indices [28] to select the final clustering results. Both the Silhouette and Dunn indices served as internal cluster validation metrics (i.e., no external information needed) to guide the selection of the final cluster set. However, the Silhouette index has been shown to have a stronger correlation with external cluster validation metrics, such as the Rand Index, than the Dun index [28,30].

The methods of Clifford et al. and Pagnuco et al. did not require selecting the number of clusters *a priori* due to the internal cluster