Phenotyping Patients with Asthma: Preprocessing, and Clustering Algorithms

Richard Peters, M.D.[1], Ali Lotfi Rezaabad[1,3]*, Matthew Sither[1], Abhishek Shende[2], and Sriram Vishwanath[3],*
[1]UT Dell Medical School, The University of Texas at Austin, [2]BrilliantMD, Inc., [3]Department of Electrical and Computer Engineering, The University of Texas at Austin,

Background. Asthma is a significant burden causing morbidity to patients and inevitable rising costs to families and the healthcare system. Current treatment is primarily based on medications to alleviate symptoms and the risk of exacerbation, but increasing emphasis is being placed on education for patients and their families, and on environmental factors. A critical issue in defining risks and appropriate diagnostics and therapeutics involves correctly classifying patients. The emergence of edge cutting techniques and developments in analytics and machine learning and the aggregation of large clinical data sets should be able to make the outcomes of machine learning models much more effective and accurate compared to traditional solutions. To this end, we study the feasibility of clustering (phenotyping) asthmatic patients based on their clinical data from electronic health records, as well as environmental and social determinants data. The intent is to pave the way for clinical interventions that will have a more targeted effect on asthma outcomes and the incidence of exacerbations. We also compared different well-known clustering algorithms to observe which can handle mixed categorical and numerical clinical data better and lead to better performance.

Methods. Clustering, the most important unsupervised learning problem, is a method partitioning n observations into k classes. Among well-known clustering algorithms are K-means, K-medoids, and density-based spatial clustering of applications with noise (DBSCAN). K-means is an iterative algorithm that tries to partition observations into k (pre-defined) non-overlapping subgroups. K-medoids is another clustering algorithm similar to K-means but the centers of a cluster is the most centered member of the cluster. On the other side, DBSCAN is a non-parametric clustering algorithm that does not require a pre-defined k, in contrast to the K-means and K-medoids algorithms. In high dimensional datasets, including clinical datasets, some features play more important roles compared to others. Accordingly, these uninformative features can be eliminated. Aside from computational benefits, this preprocessing on the dataset enables us to effectively improve the performance of clustering algorithms. For achieving these goals, we can preprocess the observations by leveraging principle component analysis (PCA) which at the same time mitigates the level of intrinsic noise in the observations. To evaluate the aforementioned methods on the asthma dataset, we compare K-means, K-medoids, and DBSCAN with and without applying PCA beforehand to study their advantages and drawbacks.

Results. We extracted the features from de-identified clinical records for asthmatic patients including patients' physical conditions of each encounter, allergies, encounter types and risk factors. We rescale each feature so they have the properties of a standard Gaussian distribution. In order to measure the performance of algorithms, we utilize silhouette score and the variance ratio criterion (VRC) which are well-known, especially as the ground truth labels are not available. Silhouette score measures how similar observations are to their associated clusters, and it ranges from -1 to +1, where high values show that observations are well matched to their associated clusters. Also, VRC is defined as the ratio between the within-cluster dispersion and between-cluster dispersion, where a higher score indicates the clusters are dense and well-separated. The results are summarized in Table 1. As it can be seen, the PCA+K-means algorithm yields better performances for the different number of clusters as compared to the other methods. It should be noted that the K-means algorithms should be initiated with the number of clusters. In the case that this is not practical, we can utilize the DBSCAN algorithms.

Table 1: Performances of the clustering algorithms for different number of clusters.

Method	Silhouette Score $(n = 3)\uparrow$	VRC($\times 10^3$) $(n = 3)\uparrow$	Silhouette Score $(n = 5)\uparrow$	VRC($\times 10^3$) $(n = 5)\uparrow$
K-means	**0.4314**	**2.515**	**0.4167**	**2.434**
K-medoids	**0.3995**	**2.458**	**0.4020**	**2.411**
DBSCAN	**0.2010**	**0.378**	**0.3380**	**0.609**
PCA + K-means	**0.4413**	**2.584**	**0.4293**	**2.551**
PCA + K-medoids	**0.4053**	**2.515**	**0.4048**	**2.392**
PCA + DBSCAN	**0.4125**	**1.071**	**0.3877**	**0.8143**

Conclusion. In this report, we study the feasibility of clustering patients with asthma. We observe that to get better results it is essential to eliminate redundant features and intrinsic noise in the dataset. Also, this study shows that the PCA+K-means algorithm yields better performance, however, it requires the number of clusters. An alternative is to leverage DBSCAN, which classifies the observations without needing to initiate the number of clusters.

*equal contributions

code is available at: https://github.com/AliLotfi92/Asthma_Study