Improved Pathogen Recognition using Non-Euclidean Distance Metrics and Weighted kNN

Mukunthan Tharmakulasingam
m.tharmakulasingam@surrey.ac.uk
Centre for Vision Speech and Signal Processing
University of Surrey
United Kingdom

Cihan Topal
c.topal@surrey.ac.uk
CVSSP, University of Surrey, UK
Eskisehir Technical University, Turkey

Anil Fernando
w.fernando@surrey.ac.uk
Centre for Vision Speech and Signal Processing
University of Surrey
United Kingdom

Roberto La Ragione
r.laragione@surrey.ac.uk
School of Veterinary Medicine
University of Surrey
United Kingdom

ABSTRACT

The timely identification of pathogens is vital in order to effectively control diseases and avoid antimicrobial resistance. Non-invasive point-of-care diagnostic tools are recently trending in identification of the pathogens and becoming a helpful tool especially for rural areas. Machine learning approaches have been widely applied on biological markers for predicting diseases and pathogens. However, there are few studies in the literature that have utilized volatile organic compounds (VOCs) as non-invasive biological markers to identify bacterial pathogens. Furthermore, there is no comprehensive study investigating the effect of different distance and similarity metrics for pathogen classification based on VOC data. In this study, we compared various non-Euclidean distance and similarity metrics with Euclidean metric to identify significantly contributing VOCs to predict pathogens. In addition, we also utilized backward feature elimination (BFE) method to accurately select the best set of features. The dataset we utilized for experiments was composed from the publications published between 1977 and 2016, and consisted of associations in between 703 VOCs and 11 pathogens. We performed extensive set of experiments with five different distance metrics in both uniform and weighted manner. Comprehensive experiments showed that it is possible to correctly predict pathogens by using 68 VOCs among 703 with 78.6% accuracy using *k*-nearest neighbour classifier and Sorensen distance metric.

CCS CONCEPTS

• **Computing methodologies** → *Classification and regression trees.*

KEYWORDS

Distance metrics, backward feature elimination (BFE), pathogen detection, bioinformatics, feature selection, machine learning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICBBE '19, November 13–15, 2019, Shanghai, China
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-7299-2/19/11…$15.00
https://doi.org/10.1145/3375923.3375956

1 INTRODUCTION

Accurate pathogen identification is essential for correct disease diagnosis, treatment of infections and identifying Antimicrobial Resistance (AMR). AMR is a growing public health threat causing 700,000 deaths per year due to AMR associated infections [1]. It is predicted to be 5 million deaths in 2050 unless urgent actions are taken to control [1]. Therefore, identifying the presence of a specific strain is the key step in preventing AMR as it allows for targeted prescribing. The proof of absence of bacterial pathogens will also help reducing unnecessary antibiotics usage.

The traditional method for identifying bacteria uses the culture of clinical specimens which is time-consuming, invasive and expensive. However, other methods can be utilized, bacteria release microbial volatile organic compounds (VOCs) in biological functions such as growth regulation, pathogenic and stress resistance. These VOCs can be used as biological markers to confirm the presence of these pathogens as humans do not produce those VOCs. Thus, combinations of VOCs representing pathogen signatures can be used as a rapid, and cheap option for the diagnosis of infectious diseases [2]. The identification of bacteria using VOCs is applied in considerable amount of clinical research due to their disease-detecting potential [3]. As numerous VOC are produced by bacteria, testing all the VOCs is not a feasible solution to identify pathogens. In addition, there is no fixed one-to-multi mapping in between pathogens and VOCs to easily identify pathogens from VOCs. For these reasons, machine learning plays a key role in identifying relevant sets of VOCs to accurately predict pathogens.

In this study, instead of utilizing subspace methods that work in metric space; we investigate the efficiency of different distance and similarity metrics on the VOC data. Our motivation was to determine if using distance metrics in other spaces other than Euclidean helps to reveal and better represent the structure of biological data, i.e. relations between VOCs and pathogens. We also perform feature selection to sort out the most meaningful VOCs