2015

Application of Supervised Machine Learning Algorithms for Rapid Identification of MRSA PFGE Strain Types

Michael L. Bernauer

University of New Mexico College of Pharmacy, bernauer@salud.unm.edu

Follow this and additional works at: https://digitalrepository.unm.edu/bmi

Part of the Bioinformatics Commons, and the Pathogenic Microbiology Commons

Recommended Citation


This Poster is brought to you for free and open access by the Health Sciences Library and Informatics Center at UNM Digital Repository. It has been accepted for inclusion in Biomedical Informatics by an authorized administrator of UNM Digital Repository. For more information, please contact disc@unm.edu.
Abstract

Background

Methods

Results

Discussion and Conclusion

References and Disclosure

Introduction: Pulsed-field gel electrophoresis (PFGE) is a molecular typing method widely used in epidemiologic surveillance of MRSA. PFGE relies on electrophoretic migration of restricted DNA fragments, resulting in characteristic fingerprints which can be used to classify organisms. Current methods of analysis, such as Gel Compare II and BioNumerics rely on unsupervised hierarchical clustering algorithms to group organisms based on pairwise similarity. These methods are labor intensive, often requiring a significant amount of user intervention and oversight. This study presents an automated approach to PFGE typing to reduce the amount of user involvement and time required for analysis.

Methods: A total of 70 gel electrophoresis images were obtained from previous PFGE experiments conducted according to protocol. Lanes were extracted from the raw images using k-means clustering, resulting in 1067 fingerprints. Transformations applied to each of the fingerprints include: normalization, alignment and background subtraction. The training set was created manually whereby each fingerprint according to PFGE type. Several fingerprints were removed due to the presence of artifacts, resulting in 916 (7% USA100, 72% USA300, 11% USA400) fingerprints for the final analysis. Principal component analysis was performed dimensionally by selecting the top 40 components. The support vector machine (SVM) was employed using 10-fold cross validation. Accuracy of the other classification algorithms was assessed by training the data 70:30 training to test. Hyperparameters for k-nearest neighbors (KNN), random forest (RF) and the artificial neural network (ANN) were selected via grid search to yield the highest test set accuracy.

Results: The ANN and SVM were the highest performing algorithms with overall classification accuracy of 0.9004, p = 1.15 × 10^-12 and 0.8845, p = 1.33 × 10^-12 with classification accuracies of 0.9283, p = 5.8 × 10^-12, 0.93 (0.920, 0.940) and 0.879, p = 0.34 × 10^-2, 0.95 (0.920, 0.960), respectively. The kNN algorithm had the greatest performance with an overall classification accuracy of 0.9158, p = 5.8 × 10^-12 and 0.879, p = 1.33 × 10^-12 with classification accuracies of 0.9283, p = 5.8 × 10^-12, 0.930 (0.920, 0.940) and 0.879, p = 0.27 (0.017, 0.960), respectively.

Conclusions: Supervised learning algorithms such as SVM and ANN provide feasible alternatives to traditional hierarchical clustering methods for determination of MRSA PFGE strain type using raw PFGE gel electrophoresis images. The aNN and SVM were the highest performing algorithms with overall classification accuracy of 0.9004, p = 1.15 × 10^-12 and 0.8845, p = 1.33 × 10^-12 with classification accuracies of 0.9283, p = 5.8 × 10^-12, 0.930 (0.920, 0.940) and 0.879, p = 0.34 × 10^-2, 0.95 (0.920, 0.960), respectively. The kNN algorithm had the greatest performance with an overall classification accuracy of 0.9158, p = 5.8 × 10^-12 and 0.879, p = 1.33 × 10^-12 with classification accuracies of 0.9283, p = 5.8 × 10^-12, 0.930 (0.920, 0.940) and 0.879, p = 0.27 (0.017, 0.960), respectively.

Limitations

• Relatively small number of training examples (n=843).
• Unbalanced dataset; USA100 (17%), USA300 (72%), USA400 (11%).
• Retrospective study design.

Authors of this presentation have no disclosures concerning possible financial or personal relationships with commercial entities that have a direct or indirect interest in the subject matter of this presentation. Authors of this presentation have no disclosures concerning possible financial or personal relationships with commercial entities that have a direct or indirect interest in the subject matter of this presentation.

References and Disclosure