World Academy of Science, Engineering and Technology
International Journal of Health and Medical Engineering
Vol:13, No:01, 2019

Experiments on Weakly-Supervised Learning on Imperfect Data

Authors : Yan Cheng, Yijun Shao, James Rudolph, Charlene R. Weir, Beth Sahlmann, Qing Zeng-Treitler

Abstract : Supervised predictive models require labeled data for training purposes. Complete and accurate labeled data, i.e., a 'gold standard', is not always available, and imperfectly labeled data may need to serve as an alternative. An important question is if the accuracy of the labeled data creates a performance ceiling for the trained model. In this study, we trained several models to recognize the presence of delirium in clinical documents using data with annotations that are not completely accurate (i.e., weakly-supervised learning). In the external evaluation, the support vector machine model with a linear kernel performed best, achieving an area under the curve of 89.3% and accuracy of 88%, surpassing the 80% accuracy of the training sample. We then generated a set of simulated data and carried out a series of experiments which demonstrated that models trained on imperfect data can (but do not always) outperform the accuracy of the training data, e.g., the area under the curve for some models is higher than 80% when trained on the data with an error rate of 40%. Our experiments also showed that the error resistance of linear modeling is associated with larger sample size, error type, and linearity of the data (all p-values < 0.001). In conclusion, this study sheds light on the usefulness of imperfect data in clinical research via weakly-supervised learning.

Keywords : weakly-supervised learning, support vector machine, prediction, delirium, simulation
Conference Title : ICHI 2019 : International Conference on Health Informatics
Conference Location : Rome, Italy
Conference Dates : January 17-18, 2019

Open Science Index, Health and Medical Engineering Vol:13, No:01, 2019 publications.waset.org/abstracts/99362.pdf