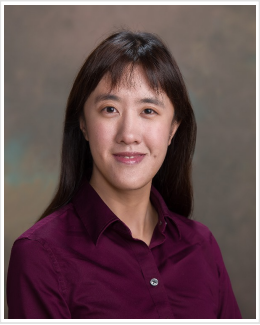
2018 Internship Project & Abstract - Zengyan Wang

Edit

An Alternative Way of Characterizing Injury to Youth on Farms



Zengyan Wang
University of Georgia

Zengyan Wang[1,2], Barbara Marlenga[2], Dick Berg[3], Peggy Peissig[2]
[1]National Farm Medicine Center, [2]Center for Computational and Biomedical Informatics, [3]Office of Research Computing and Analytics

Research area: Agricultural Health and Safety, Computational and Biomedical Informatics

Background: The incidence, types of injuries, and causes of farm-related trauma to children have been described. However, commonly used methods do not capture the complex characteristics of injuries to children on farms. We assessed complex patterns which may distinguish farm-related injuries among youth by using machine learning to conduct a high-dimensional exploration of the similarities and differences in the nature, mechanisms, and pattern of injuries to youth on farms with those occurring elsewhere.

Methods: We used data from four national public databases that capture extensive details on injury events from emergency rooms, ambulance services, trauma registries and death certificates. Due to imbalances in the databases with respect to the populations that we were comparing (farm versus non-farm injuries), both sampling and cost sensitive learning were used in the analyses. Farm injury patterns were analyzed using 1) under-sampling the majority non-farm class; 2) increasing the penalty on the misclassification on the minority farm class; 3) using random forest, extreme gradient boosting (XGBoost) and neural networks to characterize farm injuries; and 4) assessing results by area under the 'receiver operating curve' (AUC) over 10-fold cross validation.

Results: XGBoost had the best performance with a mean AUC for farm and non-farm injury classes of 0.841. Attributes derived from the XGBoost model by the importance factor increased the AUCs from 0.560 to 0.841 (compared to the dataset without those features). The important features found by the XGBoost model matched the conclusion from previous research. For instance, from the emergency rooms dataset, the top three factors related to youth injury were: 1) small hospitals in far rural areas have higher probability to accept patients; exposure to 2) machines and 3) horse-related activities are common reason to get injured.

Conclusions: Pattern discovery using massive, correlated, high-dimensional farm injury data can be effectively derived using extreme gradient boosting tree methods.