Discerning Divergent Nodes in Social Networks

Authors : Mehran Asadi, Afrand Agah

Abstract : In data mining, partitioning is used as a fundamental tool for classification. With the help of partitioning, we study the structure of data, which allows us to envision decision rules, which can be applied to classification trees. In this research, we used online social network dataset and all of its attributes (e.g., Node features, labels, etc.) to determine what constitutes an above average chance of being a divergent node. We used the R statistical computing language to conduct the analyses in this report. The data were found on the UC Irvine Machine Learning Repository. This research introduces the basic concepts of classification in online social networks. In this work, we utilize overfitting and describe different approaches for evaluation and performance comparison of different classification methods. In classification, the main objective is to categorize different items and assign them into different groups based on their properties and similarities. In data mining, recursive partitioning is being utilized to probe the structure of a data set, which allow us to envision decision rules and apply them to classify data into several groups. Estimating densities is hard, especially in high dimensions, with limited data. Of course, we do not know the densities, but we could estimate them using classical techniques. First, we calculated the correlation matrix of the dataset to see if any predictors are highly correlated with one another. By calculating the correlation coefficients for the predictor variables, we see that density is strongly correlated with transitivity. We initialized a data frame to easily compare the quality of the result classification methods and utilized decision trees (with k-fold cross validation to prune the tree). The method performed on this dataset is decision trees. Decision tree is a non-parametric classification method, which uses a set of rules to predict that each observation belongs to the most commonly occurring class label of the training data. Our method aggregates many decision trees to create an optimized model that is not susceptible to overfitting. When using a decision tree, however, it is important to use cross-validation to prune the tree in order to narrow it down to the most important variables.

Keywords : online social networks, data mining, social cloud computing, interaction and collaboration
Conference Title : ICSSR 2020 : International Conference on Social Computing and Applications
Conference Location : Singapore, Singapore
Conference Dates : September 08-09, 2020