CS36110 Assignment: Employee Absenteeism

Ben Jaeger | bej29

Contents

[1. Initial Building of Machine Learning Models 3](#_Toc23417331)

[1.1. Initial Classifier Creation and Explanation 3](#_Toc23417332)

[1.2. Comparing Results between Classifiers 3](#_Toc23417333)

[References 3](#_Toc23417334)

# 1. Initial Building of Machine Learning Models

## 1.1. Initial Classifier Creation and Explanation

For this assignment I used a J48 and NaiveBayes classifier as prescribed in the assignment, however I also used a RandomForest classifier as I had a passing knowledge of it from previous machine learning work.

J48 is an implementation of a decision tree-based classifier based on the C4.5 release 8 classifier (hence the name). It works by finding the entropy within a dataset, this defines how heterogeneous the dataset is, seemingly unrelated the features are to the decision feature, then it defines the gain that would be made by dividing the dataset across a feature, the gain being a measure of the reduction in entropy. The aim being to greedily reduce entropy, and thus increase the homogeneity at each level. One advantage of J48 is that it is a very simple model to both produce and, arguably more importantly in some situations, a explain to someone who has little or no knowledge of machine learning.

NaiveBayes is a probabilistic classifier, meaning it uses the given dataset to find the probability of an instance belonging to a particular decision feature group. This is done by trying to calculate how likely the outcome is given the features (“how likely is it a banana given it’s long, yellow and sweet”), this is achieved by finding how likely that outcome is in the first place (“how likely is it that this is a banana”) the logic being that less likely outcomes should be prejudiced against and also how likely each of those features is given that outcome (“if this were a banana, how likely is it that its long yellow and sweet”) this aims to prejudice against unlikely combinations resulting in that outcome. NaiveBayes has the advantage of disregarding the interactions between features (hence why its naive) this reduces the calculations needed greatly. It also easily adapts to more data as there is nothing to restructure such as a decision tree.

The last classifier is a RandomForest classifier as I had encountered it before. A RandomForest is an example of an ensemble learning model, meaning that it is a model made up of many smaller models, attempting to negate the negatives of each. The classifier takes many decision trees formed from subsets of the dataset (with replacement) and then “votes” between them, the idea being that this will reduce the chance of the tree over fitting to the data as just one decision tree, such as J48, is sometimes susceptible to doing. The difference between a normal decision tree and those that make up this model is that at each level the tree randomly pick from a subset of the features remaining, when it comes to voting the majority rules at each level for classification (Liaw, 2002).

## 1.2. Comparing Results between Classifiers

When comparing machine learning models there are many ways to measure and compare them. We will be using the percentage correct as well as the ROC area,

According to WEKA 3.8 (the tool used to create these models) the NaiveBayes classifier had the lowest percentage correct at 55.69% with J48 at 71.935% and RandomForest narrowly coming out on top with 78.582%

###

# References

Liaw, A. &. ( 2002). Classification and regression by randomForest. *R news*, 2(3) 18-22.