doi:10.20396/revpibic.v0i0.id

AN INTRODUCTORY STUDY TO MACHINE LEARNING AND ITS APPLICATION TO EMPLOYEE TURNOVER PREDICTION

João Pedro Pazinato Cruz de Oliveira*

Abstract
The objective of this paper is to study the problem of employee turnover prediction and to develop a classifier that uses employee's data to identify those who have a greater tendency to leave the company voluntarily. For such purpose, the data of 8724 employees from a real Brazilian beverage company was used to train an Extreme Learning Machine (ELM) classifier, assigning to each sample a weight inversely proportional to the size of the respective class. After the training, the classifier displayed an overall accuracy of 79% of the test data.

Key words:
Machine Learning, Turnover, Extreme Learning Machines

Introduction

Employee turnover can cause instability and unexpected expenses for the organization. Therefore, it is desirable that a company be able to accurately identify the employee's intention to leave and take appropriate action to decrease its turnover rate.

The motivations leading an employee to leave a company are often complex and difficult to identify, so machine learning models have been used successfully to address problems similar to the Turnover problem.

In supervised learning algorithms, the model goes through a training step that is performed automatically from existing data, after the training stage the model is able to make predictions from new data.

The present work studied several machine learning models and applied the knowledge studied using Extreme Learning Machines (ELM), a type of artificial neural network with a single hidden layer whose weights are fixed randomly and the weights of the output layer can be trained through the least squares method, which grants ELMs a simple and fast training, alongside a good generalization performance [1].

Results and Discussion

The data of 8724 employees of a real Brazilian beverage company were used. 15% of them voluntarily resigned, 11% were dismissed from their jobs, and 74% remain active in the company.

The database contains information such as employee's age, marital status, number of children, wage, hierarchy level, job tenure, job satisfaction, goals achievement, among others. In total, 20 independent variables were used.

The database was randomly split so that 70% of the data were used for the training step and 30% of the data were used to test the performance of the model after training.

During the training stage, the classes of employees who were involuntarily dismissed and those who remained active in the company were grouped as one, so that classifier outputs a number between zero and one, indicating a lower or greater probability of that employee resigning voluntarily.

Given the low proportion of people who resigned voluntarily in the database, a class balancing method was used applying a weight inversely proportional to the size of the sample's class to each sample.

Figures 1, 2 and 3 show the receiver operating characteristic (ROC) curve, the confusion matrix for a threshold = 0.5, and the histogram of the ELM output for each class, all of which are generated from the test data. The overall accuracy of the model was 79%.

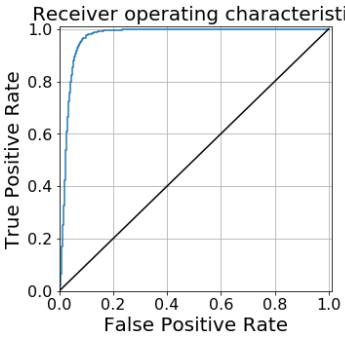

Figure 1. ROC curve.

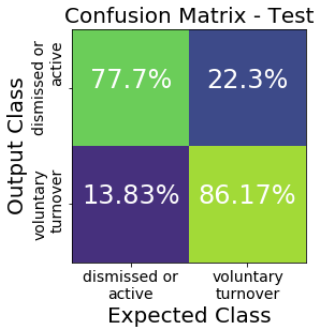

Figure 2. Confusion matrix.

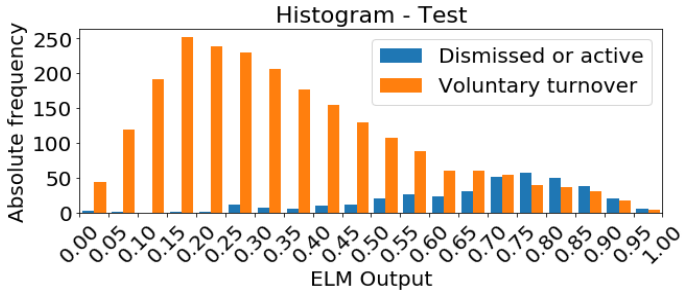

Figure 3. Histogram of the ELM output.

Conclusions

The ELM and class balancing method presented an adequate performance for the turnover data used, with an overall accuracy of 79% of the test data and an accuracy of 86% for the voluntary turnover class, fulfilling their purpose of indicating which employees would have a greater chance of resigning.

[1] HUANG, Guang-Bin et al. Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), v. 42, n. 2, p. 513-529, 2012.