# Penerapan bayes risk post pruning pada klasifikasi decision tree = Application of bayes risk post pruning in decision tree classification

Devina Christianti, author



---

Abstrak

Klasifikasi adalah proses menugaskan satu set data ke dalam kelas yang ada berdasarkan nilai setiap atribut. Pengklasifikasi pohon keputusan diklaim lebih cepat dan berproduksi akurasi yang lebih baik. Namun, ia memiliki beberapa kelemahan di mana pengklasifikasi rentan untuk overfitting. Overfitting adalah suatu kondisi di mana model tidak mampu menarik kesimpulan data baru dengan cara yang benar. Overfitting di pohon keputusan dapat dihindari dengan memotong subtree pengaruh kecil dalam melakukan klasifikasi ketika pohon ditanam, disebut post-pruning, yang bertujuan untuk meningkatkan kinerja model dalam memprediksi data.
Tesis ini mengusulkan metode pasca pemangkasan dengan menerapkan Risiko Bayes, di mana estimasi risiko setiap simpul induk dibandingkan dengan simpul daunnya. Sebagai perbandingan, pemangkasan pasca lainnya Metode yang diterapkan, yaitu Reduced Error Pruning (REP). Kedua metode tersebut diterapkan untuk tiga dataset klasifikasi churn pelanggan dari situs Kaggle dan IBM Datasets. Untuk hasilnya, Bayes Risk Post-Pruning dapat meningkatkan kinerja Decision Tree lebih baik dari Reduced Error Pruning dengan meningkatkan nilai akurasi, presisi, dan daya ingat. Kedua metode juga diterapkan pada tiga proporsi berbeda untuk data pelatihan (60%, 70% dan 80%). Hasilnya menunjukkan bahwa semakin besar ukuran dataset pelatihan dikaitkan akurasi, presisi, dan daya ingat model yang lebih tinggi.
.....
Classification is the process of assigning a set of data to an existing class based on the value of each attribute. Decision tree classifiers are claimed to be faster and produce better accuracy. However, it has several disadvantages where the classifier is prone to overfitting. Overfitting is a condition in which the model is unable to draw new data conclusions in the right way. Overfitting in the decision tree can be avoided by cutting the subtree of small influence in classifying when the tree is planted, called post-pruning, which aims to improve the performance of the model in predicting data.
This thesis propose a post-pruning method by applying Bayes Risk, where the estimated risk of each parent node is compared to the leaf node. As a comparison, other post pruning methods are applied, namely Reduced Error Pruning (REP). Both methods are applied for three customer churn classification datasets from the Kaggle site and IBM Datasets. For the results, Bayes Risk Post-Pruning can improve Decision Tree performance better than Reduced Error Pruning by increasing the value of accuracy, precision, and memory. Both methods are also applied to three different proportions for training data (60%, 70% and 80%). The results show that the greater the size of the training dataset is associated with higher model accuracy, precision, and recall.