Jurnal Ilmiah Ilmu Komputer, Vol 15 No. 2, Desember 2010 : 7 - 13

Perbandingan Algoritme *Pruning* pada *Decision Tree* yang Dikembangkan dengan Algoritme CART

Martin Budi, Rindang Karyadin, Sony Hartono Wijaya

Departemen Ilmu Komputer, Institut Pertanian Bogor, Jl. Meranti Wing 20 Lv.V, Bogor, Jawa Barat, 16680

Abstract---Pruning is part of the development of decision tree. As decision tree is developed, some nodes became outliers as the results of noise data. Implementation of the decision tree pruning, can reduce the noise and outliers on the initial decision tree so that it can improve the accuracy of the data classification. Therefore the selection of proper pruning algorithm needs to be done to get the maximum results of the classification.

This experiment uses data from the company's customer credit providers. The data obtained from the data bank at the University of California. Data used in this experiment has twenty variables with two classes and 1000 instances. The data contain thirteen qualitative variables and the rest is a numeric data. The data is a good for use because it does not have a missing value.

The experiment compared three pruning algorithm, Cost Complexity Pruning (CCP), Reduced Error Pruning (REP), Error Based Pruning (EBP). Those pruning algorithms do prune to the decision tree that was developed with the Classification and Regression Tree (CART) algorithm. The comparison of those algorithms is done repeatedly on the data with different conditions both in terms of the instance number and the data variables. Comparison of the algorithm includes a comparison of the accuracy of the decision tree, and the process time of pruning algorithm.

The experiment's result shows the average error rate of that the REP algorithm will produce the smallest error rate. Although the error rate of REP algorithm is the smallest, the difference value between ERP's and EBP's error rate is only 0.5%. Even though they have almost similar error rate, EBP algorithm proposes more simple decision tree than REP algorithm does.

Keyword : Decision tree, Classification and Regression Tree (CART), Cost Complexity Pruning (CCP), Reduced Error Pruning (REP), Error Based Pruning (EBP).

PENDAHULUAN

Data mining merupakan salah satu tahapan dalam proses *Knowledge Discovery in Database* (KDD) yang melakukan ekstraksi informasi atau pola penting dalam data berukuran besar (Han & Kamber 2006). Teknik yang dapat digunakan pada implementasi *data mining* adalah klasifikasi dan prediksi, *association rule*, dan *clustering*. Klasifikasi merupakan metode yang berfungsi untuk menemukan model yang membedakan kelas data, sehingga klasifikasi dapat memperkirakan label kelas dari suatu objek yang belum diketahui. Salah satu metode klasifikasi yang sering digunakan adalah *decision tree*.

Pruning merupakan bagian dari proses pembentukan *decision tree*. Saat pembentukan *decision tree*, beberapa *node* merupakan
outlier maupun hasil dari *noise* data. Penerapan *pruning* pada *decision tree*, dapat mengurangi *outlier* maupun *noise* data pada *decision tree* awal sehingga dapat meningkatkan akurasi pada klasifikasi data (Han & Kamber 2006). Oleh sebab itu pemilihan algoritme *pruning* yang tepat perlu dilakukan untuk mendapat hasil klasifikasi yang maksimal.

Pada penelitian yang dilakukan oleh (Esposito *et al.* 1997), algoritme Reduced Error Pruning (REP) disimpulkan sebagai algoritme yang menghasilkan *subtree* terkecil dengan *error rate* minimum. Penelitian Esposito *et al.* 1997) menggunakan algoritme C4.5 untuk membangun *decision tree* yang di-*pruning*. Berbeda dengan penelitian sebelumnya, penelitian ini membandingkan pengunaan algoritme *pruning* pada *decision tree* yang dibangun dengan algoritme Classification and Regression Tree (CART). Algoritme CART biasa menggunakan Cost Complexity Pruning (CCP) sebagai algoritme *pruning*-nya. Pada penelitian ini algoritme *pruning* CCP dibandingkan dengan dua algortime *pruning* lain yaitu REP dan Error Based Pruning (EBP).

Penelitian in bertujuan untuk menerapkan teknik CCP , REP dan EBP pada metode klasifikasi *decision tree* dengan algoritme CART. Selain itu penelitian in ijuga membandingkan nilai akurasi dari *decision tree* yang terbentuk, serta waktu proses yang dihasilkan oleh algoritme *pruning* CCP , REP dan EBP.

METODE PENELITIAN

Penelitian ini menerapkan tahapan yang tertuang dalam suatu Metodologi Penelitian (Gambar 1).

A. Studi Literatur

Studi literatur dilakukan dengan memperdalam algoritme-algoritme yang akan dibandingkan. Informasi yang diperoleh berasal dari beberapa sumber seperti : jurnal, buku dan artikel di internet.

B. Pengumpulan Data

Penelitian ini menggunakan data *profile* pelanggan dari perusahaan penyedia kredit. Data tersebut diperoleh dari bank data pada University of California (Asuncion & Newman 2007). Contoh data dapat dilihat pada Lampiran

C. *Data Mining*

Teknik *data mining* dengan metode *decision tree* terdiri dari dua tahapan, yaitu: