ABSTRAK

Algoritme pohon keputusan (*decision tree*) C4.5 merupakan salah satu algoritme klasifikasi data yang mudah untuk diintepretasikan dan mempunyai struktur yang sederhana. Beberapa penelitian menjadikan *decision tree* sebagai objek kajian, yaitu pada proses pemilihan *split* atribut dan proses *pruning*. Pemilihan *Split* atribut merupakan proses utama dalam pembentukan *decision tree*. Metode-metode yang dikembangkan untuk pemilihan *split* atribut diantaranya *Gini Index, Information Gain, Simplifying Decision Tree, Gain Ratio, Imprecise Info Gain* (IIG), *Imprecise Information Gain Ratio* (IIGR), dan *AdaptiveCC4.5.* Setelah dilakukan pemilihan *split* atribut, proses selanjutnya adalah *pruning*. *Pruning* adalah proses yang dilakukan untuk memotong atau menghilangkan beberapa cabang (*node*) yang tidak diperlukan. *Node* yang tidak diperlukan dapat menyebabkan ukuran *decision tree* menjadi sangat besar dan hal ini disebut *over-fitting.* Metode-metode yang dikembangkan untuk *pruning* diantaranya menggunakan *post pruning*, yaitu *Pessimistic Error Pruning, Cost complexity pruning* (CCP), *Reduce Error Pruning* (REP), dan *Error Base Pruning* (EBP). Beberapa penelitian yang dikembangkan untuk proses *pruning* masih memiliki kelemahan, yaitu memungkinkan pemangkasan terhadap *node* dengan nilai informasi yang tinggi dan kontribusi dalam pembentukan *rule*. Penelitian ini mengusulkan metode untuk memodifikasi proses *pruning* menggunakan perhitungan nilai delta, *certainty factor* (CF) dan *weighting value*. Nilai delta ini diterapkan dengan memanfaatkan perbedaan jumlah anggota objek antar *node*. Apabila perbedaan jumlah anggota objek antar *node* (selisih) kecil menunjukkan bahwa *node* tersebut adalah *node* yang tidak kontributif. Nilai CF diperoleh dari nilai entropi yang dikalikan dengan selisih dari entropi maksimum dengan nilai *Gain_Ratio*. Perhitungan nilai CF digunakan untuk memastikan bahwa *node* yang akan dilakukan proses *pruning* adalah *node* yang benar-benar tidak kontributif. Sedangkan *weighting value* digunakan untuk memperbaiki proses *pruning* dengan memberikan nilai bobot pada setiap atribut yang terpilih sebagai *Split*. Nilai bobot ini akan membatasi pohon keputusan yang dibentuk. Pengembangan yang dilakukan pada penelitian ini adalah dengan memanfaatkan *resources* yang sudah ada pada C4.5, yaitu dengan menggunakan fungsi entropi untuk menentukan cabang (*node*) yang akan di-*pruning*, sehingga tambahan beban komputasi sangat kecil, yaitu pada proses pemangkasannya saja. *Dataset* yang digunakan untuk eksperimen dan menjelaskan metode peningkatan nilai akurasi Algoritme C4.5 ini adalah *dataset* dari *University of California at Irvine's* (UCI) *machine learning repository*. Setelah dilakukan pengukuran nilai akurasi terhadap seluruh *dataset* yang digunakan, kemudian dilakukan perbandingan nilai akurasi yang dihasilkan oleh masing-masing metode. Hasil eksperimen pada penelitian ini menunjukkan bahwa C4.5+*Weighting value* memiliki performa yang paling baik dibandingkan dengan metode lainnya. Hal ini ditunjukkan dengan peningkatan nilai akurasi sebesar 8,48% (C4.5+REP), 2,87% (CF), dan 1,9% (Delta). Hasil dari penelitian ini dapat digunakan sebagai acuan untuk memilih metode yang tepat sesuai dengan kasus-kasus dalam data mining khususnya proses klasifikasi data.

Kata kunci: peningkatan nilai akurasi Algoritme C4.5, *pruning*, cabang kontributif

ABSTRACT

C4.5 algorithm is one of the decision tree algorithms that can be used to generate rules that are easily interpreted and fastest among other algorithms. The algorithm is also able to produce a model in a support decision support systems, although it still requires a performance improvement. There are many features involved in C4.5 Algorithm, namely: data, data attributes, instances, and attribute classes. This algorithm still produces a poor level of accuracy in some classification cases, especially in the pruning process. Several methods that have been developed in previous studies have provided increased accuracy, but still raise the pruning problem of pruning with high information values. More pruning is done to nodes with high information values, it will cause a decrease in the level of accuracy. This study raises the pruning error process problem on contributive nodes. The purpose of this study is to improve C4.5 algorithm performance by pruning process improvement. This problem was solved by proposing pruning improvements using three methods, which are C4.5 + Delta, C4.5 + CF, and C4.5 + Weighting value. Overall, the contribution of this research is pruning process improvement on decision tree algorithms which will improve decision tree and its derived algorithms namely: C4.5 + delta algorithm, C4.5 + CF, and C4 algorithms. It also makes ease for researchers to use the modified C4.5 of this study to be further developed or combined with other methods. Dataset used for the experiment and explains performance improvement of C4.5 algorithm is a dataset from the University of California at Irvine's (UCI) machine learning repository. After measuring the accuracy value of all datasets used, then an increase in the average value of accuracy is calculated. Experimental results in this study indicate that C4.5 + Weighting value has the best performance compared to other methods. This is indicated by an increase in the average accuracy of 8.48% (C4.5 + REP), 2.87% (CF), and 1.9% (Delta)

Keywords—pruning, contributive nodes, C4.5 Algorithm