INTISARI

PENINGKATAN AKURASI ALGORITMA C4.5 MENGGUNAKAN DISKRITISASI DAN PEMILIHAN FITUR BERBASIS KORELASI

Oleh

Gita Ayu Wulan Sari

17/409518/PA/17825

Salah satu teknik dari *supervised learning* pada *machine learning* adalah klasifikasi, dan salah satu algoritma klasifikasi adalah C4.5. Tujuan dari penelitian ini adalah untuk meningkatkan akurasi algoritma C4.5 dengan menerapkan diskritisasi dan *Correlation-Based Feature Selection* (CFS). Akurasi yang lebih baik telah dicapai dengan menerapkan diskritisasi dan CFS. Diskritisasi digunakan untuk menangani nilai kontinu, sedangkan CFS digunakan sebagai pemilihan atribut.

Di bidang kesehatan, teknik klasifikasi dapat digunakan untuk mendiagnosis penyakit dari data rekam medis pasien. Penelitian ini menggunakan empat *dataset* yang diperoleh dari UCI *Machine Learning Repository*. Dalam *dataset* ini, terdiri dari atribut berupa tipe numerik yang kontinu dan nominal. Atribut kontinu dapat menyebabkan akurasi yang rendah karena bentuk data yang tidak terbatas. Sedemikian hingga, atribut perlu diubah menjadi data diskrit dengan diskritisasi. Selain itu, pada kasus tertentu jika semua atribut digunakan dapat menghasilkan tingkat akurasi yang rendah karena tidak relevan dan tidak memiliki korelasi dengan kelas target. Oleh karena itu, setelah didiskritisasi atribut tersebut perlu diseleksi terlebih dahulu untuk mendapatkan hasil yang lebih akurat menggunakan CFS.

Kata Kunci: Klasifikasi, Algoritma C4.5, Diskritisasi, Entropi, *Minimal Description Length*, *Correlation-Based Feature Selection*.

ABSTRACT

INCREASING ACCURACY OF C.45 ALGORITHM BY USING DISCRETIZATION AND CORRELATION-BASED FEATURE SELECTION

By

Gita Ayu Wulan Sari

17/409518/PA/17825

One of the techniques of supervised learning in machine learning is classification, and one of the classification algorithms is C4.5. The purpose of this study is to increase the accuracy of the C4.5 algorithm by using discretization and Correlation-Based Feature Selection (CFS). Better accuracy has been achieved by using discretization and CFS. Discretization is used to handle continuous values, while CFS is used as attribute selection.

In the health sector, classification can be used to diagnose diseases from patient medical records. This study uses four datasets obtained from the UCI Machine Learning Repository. In this dataset, it consists of attributes in the form of numeric and nominal types. The continuous attribute can cause low accuracy due to the infinite form of the data. Thus, attributes need to be converted into discrete data with discretization. In addition, in certain cases if all attributes are used it can result in a low level of accuracy because it is irrelevant and has no correlation with the target class. Therefore, after being discretized these attributes need to be selected first to get more accurate results using CFS method.

***Keywords**: Classification, C4.5 Algorithm, Discretization, Entropy, Minimal Description Length, Correlation-Based Feature Selection.*