

ABSTRAK

Penambangan data adalah proses mengekstrak informasi atau pengetahuan dari data dalam jumlah yang besar. Secara umum, penambangan data dapat diklasifikasikan dalam empat kategori yaitu deteksi dependensi, identifikasi kelas, deskripsi kelas dan deteksi *outlier*. Tiga kategori pertama berkaitan dengan pola yang dimiliki oleh banyak objek, atau pada objek dengan persentase yang besar dalam *dataset*. Sebaliknya, kategori keempat berfokus pada objek dengan persentase yang kecil, yang umumnya sering diabaikan atau dihilangkan karena dianggap *noise* (Knorr & Ng, 1998).

Pada tugas akhir ini algoritma *Block-based Nested-Loop* digunakan untuk deteksi *outlier* pada data numerik. Data yang digunakan adalah data akademik mahasiswa Universitas Sanata Dharma Prodi Teknik Informatika Angkatan 2007 dan 2008. Data tersebut terdiri dari data nilai tes masuk mahasiswa yang diterima melalui jalur tes tertulis maupun jalur prestasi dan nilai indeks prestasi dari semester satu sampai empat.

Hasil penelitian ini yaitu sebuah perangkat lunak yang dapat digunakan sebagai alat bantu untuk mendeteksi *outlier* menggunakan algoritma *Block-based Nested-Loop*. Pengujian terhadap sistem ini terdiri dari pengujian *blackbox*, pengujian *review* dan validitas oleh pengguna, pengujian efek perubahan nilai atribut penambangan data dan pengujian waktu deteksi *outlier* berdasarkan jumlah blok.

Berdasarkan hasil pengujian *blackbox* yang dilakukan, disimpulkan bahwa sistem pendeteksi *outlier* ini dapat menghasilkan keluaran yang sesuai dengan yang diharapkan pengguna dan mampu menangani *error* terhadap fungsi – fungsi yang tidak berjalan sesuai aturan. Berdasarkan hasil pengujian *review* dan validitas oleh pengguna disimpulkan bahwa sistem dapat menghasilkan data yang dinyatakan sebagai *outlier*. Berdasarkan hasil pengujian efek perubahan nilai atribut penambangan data disimpulkan bahwa pemilihan nilai atribut M dan D mempengaruhi hasil deteksi *outlier*. Dari hasil pengujian waktu deteksi *outlier* berdasarkan jumlah blok disimpulkan bahwa penggunaan blok-blok data dapat mempercepat proses deteksi *outlier*.

Kata kunci : Penambangan data, *Block-based Nested-Loop*, deteksi *outlier*

ABSTRACT

Data mining is the process of extracting information or knowledge from large amounts of data. In general, data mining can be classified into four categories: dependency detection, identification of classes, class descriptions and outlier detection. The first three categories correspond to the pattern that applied to many objects, or to a large percentage of objects in the dataset. The fourth category, in contrast focuses on a very small percentage of data objects, which are often ignored or discarded as noise (Knorr & Ng, 1998).

In this thesis, the Block-based Nested-Loop algorithm was used to perform outlier detections on numerical data. The data used in this thesis are academic data of students batch 2007 and 2008 of Informatics Engineering Study Program of Sanata Dharma University. The data consists of student admission data from regular admission track as well as students from outstanding track, and student academic data (Grade Point Average) of those students from first semester until fourth semester.

The results of this research is a software that can be used as a tool to detect outliers using Block-based Nested-Loop algorithm. The testing of this system consists of blackbox testing, validation testing by users, investigation of the effects of attribute changes of M and D, and investigation of time needed to detect outliers based on the number of blocks. M is the maximum number of objects within the D-neighbourhood of an outlier, whereas D is the maximum distance between any pair of objects that define as a neighbor.

Based on blackbox testing, it can be concluded that the outlier detection's system could produce output as expected and handle any incorrect functions. Based on user's validation, it can be concluded that the results of the system are confirmed as outliers. Based on the investigation of the effects of attribute changes, it can be concluded that the value of M dan D influence the number of generated outliers. Based on investigation of time needed to detect outliers, it can be concluded that data blocks usage could speed up the process of outlier detection.

Keywords : *data mining, Block-based Nested-Loop, outlier detection*