PLAGIAT MERUPAKAN TINDAKAN TIDAK TERPUJI

ABSTRAK

Outlier adalah adalah obyek yang berbeda dibandingkan obyek – obyek lain dalam suatu dataset. Dalam penambangan data, deteksi *outlier* adalah satu satu bidang penelitian yang terus berkembang. Umumnya metode deteksi *outlier* tidak memperhatikan secara khusus *class label* pada dataset dan hanya fokus pada dataset yang seragam. Padahal, dataset yang nyata biasanya mempunyai multiatribut. Pada deteksi *outlier* dengan algoritma *Enhanced Class Outlier Distance Based* (ECODB), data yang menyimpang dari kumpulan *class*-nya dapat ditemukan. Algoritma ECODB dapat diterapkan pada dataset dengan atribut campuran numerik dan kategorikal.

Algoritma ECODB akan menghitung nilai *Class Outlier Factor* (COF) dari tiap *instances* berdasarkan masukan nilai k dan *top N*. K adalah jumlah tetangga terdekat dari suatu *instances*, sedangkan *top N* adalah jumlah *instances* yang dideteksi sebagai *outlier* yang diurutkan secara kecil ke besar berdasarkan nilai COF. COF adalah nilai probabilitas/derajat sebuah *instance* dapat menjadi *outlier*. *Outlier* adalah data dengan nilai COF terendah.

Pada penelitian ini dilakukan pendeteksian *outlier* menggunakan algoritma ECODB. Data yang digunakan adalah data debitur BPR XYZ yang mengangsur kredit pada bulan Agustus 2013. Data tersebut berjumlah 97 *record* dalam format Microsoft Excel (.xls). Pada penelitian ini akan diketahui bagaimana pengaruh nilai k dan *top N* dalam proses deteksi *outlier* menggunakan algoritma ECODB.

Pengujian dilakukan dengan cara menghitung data debitur BPR XYZ menggunakan algoritma ECODB dengan masukan k dan *top N* yang berbeda. Kemudian hasil perhitungan tersebut dibandingkan untuk mendapatkan kesimpulan. Selain itu juga dilakukan review hasil deteksi *outlier* oleh petugas bank.

Dari hasil pengujian efek perubahan nilai k dan *top N* dapat disimpulkan bahwa penentuan nilai k dan *top N* pada algoritma ECODB berpengaruh terhadap *outlier* yang dihasilkan. Nilai k dan top N yang terlalu kecil atau besar menyebabkan hasil deteksi outlier tidak optimal. Berdasarkan hasil pengujian *review* dan validitas oleh petugas bank dapat disimpulkan bahwa hasil deteksi *outlier* yang diperoleh layak dinyatakan sebagai *outlier*.

Kata kunci : penambangan data, deteksi *outlier*, ecodb, *enhanced class outlier distance based*

PLAGIAT MERUPAKAN TINDAKAN TIDAK TERPUJI

ABSTRACT

Outlier is an object which is different from any objects in one dataset. In data mining, outlier detection is one of growing researches. Generally, outlier detection methods find exception or rare cases in a dataset without considered class label as an important thing and only can be used on dataset that have single datatypes. In fact, real world dataset usually have mixed datatypes. On outlier detection using Enhanced Class Outlier Distance Based (ECODB) algorithm, data which is different from its class can be found. ECODB algorithm can be applied on dataset that have numerical and categorical attributes.

ECODB algorithm count the Class Outlier Factor (COF) from each instances based on k and top N value. K is the nearest neighbors of instances, whereas top N is the number of top class outlier that rank from greatest to the least based on COF value. COF is the probability/degree from an instance to be considered as outlier. Outlier is data which have least COF value.

In this thesis, ECODB algorithm was used to perform outlier detection. The data used in this thesis is credit data of BPR XYZ debtor whom lessened their credit on August 2013. This data consist of 97 records on Microsoft Excel format (.xls). In this thesis, it can be understand how k and top N value influenced on outlier detection using ECODB algorithm.

The testing can be done by counting credit data of BPR XYZ using ECODB algoritm with various input of k and top N. The results was compared to provide the conclusion. Besides, it also validated the results of outlier detection by reviewing the bank officer.

Based on the testing, it can be concluded that the determination of k and top N value influence the results of outlier detection. Very small or very high of k and top N value cause unoptimal outlier detection. Also, based on validation testing by bank officer, the results of the outlier detection using ECODB algorithm are confirmed as outliers.

Keyword : data mining, outlier detection, ecodb, enhanced class outlier distance based