INTISARI

PERBANDINGAN ALGORITMA METODE DETEKSI OUTLIER PADA DATA KATEGORIK

Oleh

Khoirul Umam Marjianto

14/373528/PA/16411

Era perkembangan teknologi informasi saat ini, memungkinkan untuk menyimpan data dalam jumlah besar. *Data mining* merupakan sebuah disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan suatu pola tertentu dari suatu kumpulan data yang besar. Pada *data mining*, data dengan karakteristik yang berbeda dari data pada umumnya serta mempunyai kemunculan yang relatif sedikit disebut sebagai sebuah *outlier*. Pada saat ini telah banyak dikembangkan metode deteksi *outlier*, namun umumnya untuk data numerik. Pada penelitian ini akan dibahas perbandingan metode deteksi *outlier* pada data kategorik. Adapun algoritma yang akan dibandingkan pada penelitian ini adalah AVF, NAVF, dan AEVF. Pemilihan ketiga algoritma tersebut karena masing-masing algoritma tersebut unggul pada saat dilakukan pengujian pada penelitian yang sudah dilakukan sebelumnya.

Pada penelitian ini akan dilakukan analisis perbandingan antara algoritma AVF, NAVF, dan AEVF, meliputi *detection rate* dan waktu tempuh algoritma. Pada perbandingan *detection rate*, dataset akan dibagi menjadi data normal dan *outlier* kemudian data tersebut digabungkan dan dilakukan deteksi *outlier*. Sedangkan pada perbandingan waktu tempuh, dilakukan perhitungan waktu berdasarkan perubahan dimensi dataset, jumlah atribut dan perbedaan nilai atribut.

Hasil dari penelitian ini, algoritma AVF memberikan akurasi dan waktu yang lebih baik dibanding NAVF dan AEVF. Namun AVF mempunyai kelemahan yaitu harus memasukan k *outlier*, alternatif solusinya adalah menggunakan NAVF yang mempunyai akurasi dan waktu tempuh yang berbeda sedikit dibawah AVF namun tidak memerlukan masukan k *outlier*. AVF dan NAVF cocok digunakan pada dataset dengan jumlah atribut dan perbedaan nilai atribut yang sedikit serta persebaran data yang merata. Sedangkan AEVF cocok untuk digunakan pada kondisi sebaliknya. Hal yang mempengaruhi waktu eksekusi algoritma AVF, NAVF dan AEVF adalah perubahan jumlah dimensi data, jumlah atribut data, dan jumlah nilai pada atribut data.

Kata kunci : deteksi *outlier*, data kategorik, AVF, NAVF, AEVF, *data mining*, *pre-processing data*, *detection rate*, waktu tempuh

ABSTRACT

COMPARISON OF OUTLIER DETECTION ALGORITHM METHOD ON CATEGORICAL DATA

By

Khoirul Umam Marjianto

14/373528/PA/16411

In the development of information technology today, it is possible to store large amounts of data. Data mining is a discipline that studies the method to extract knowledge or find a particular pattern of a large data set. In the data mining, the data with different characteristics and have little appearance known as outliers. At this time has been developed outlier detection methods, but generally for numerical data This research will be discussed outlier detection method comparison on categorical data as for the algorithm to be compared in this study are AVF, NAVF, and AEVF. The third Selection of these algorithm is because each algorithm has a good result in research that has been done before.

This research will perform a comparative analysis between AVF, NAVF and AEVF algorithms, includes the detection rate and the running time algorithm. In the comparison of detection rate, the dataset will be divided into normal data and outliers then the data is combined and performed outlier detection. Then on the comparison running time, the calculation will be done based on change of dimensional dataset, number of attributes, and distict attribute value per attribute.

The results of this research is AVF algorithm provides outlier detection accuracy and a better time than NAVF and AEVF. But AVF has the disadvantage that must include k outlier, an alternative solution is to use NAVF which has the accuracy and the running time is a little different under AVF but don't require input of k outlier. AVF and NAVF algorithms are suitable for use in a dataset which the number of attributes and distinct attribute values per attribute are little differences and also uneven distribution of data. While AEVF suitable for use in conditions contrary AVF and NAVF. The thing that affects the algorithm running time AVF, NAVF and AEVF are, changes of the number of data dimensions, changes of number attributes, and change of distict attribute value per attribute.

Keywords: outlier detection, categorical data, AVF, NAVF, AEVF Greedy, data mining, pre-processing data, detection rate, running time