DETEKSI OUTLIER PADA CATEGORICAL DATA MENGGUNAKAN ALGORITMA LSA (LOCAL SEARCH ALGORITHM)

Aditya Pamungkas[1], Kiki Maulana[2], Angelina Prima Kurniati[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Data Mining adalah proses pencarian pola-pola dan kecenderungan yang menarik dari dalam basis data berukuran besar. Sebuah outlier didefinisikan sebagai sebuah titik data pada suatu data set dimana sangat berbeda dibandingkan dengan titik data pada data set pada umumnya dengan suatu ukuran tertentu. Outlier ini walaupun mempunyai kelakuan yang abnormal, seringkali mengandung informasi yang sangat berguna.
Fungsi dari deteksi outlier adalah untuk mencari sekelompok kecil objek data yang merupakan pengecualian ketika dibandingkan dengan sejumlah besar data yang lainnya. Deteksi Outlier seperti ini sangat penting untuk banyak aplikasi seperti fraud detection, intrusion detection, dan network monitoring application. Kebanyakan metoda yang ada didesain untuk data numerik. Ini akan mengatasi masalah yang ada pada aplikasi di dunia nyata yang mengandung data kategoris. Metoda LSA dapat mendeteksi outlier dengan menggunakan local-search heuristik dan rumus entropy untuk perhitungannya.

Kata Kunci : data mining, outlier, deteksi outlier, LSA, entropy

Abstract

Data mining is interesting patterns and trend finding process in large database. Outlier defined as a data point in database where is different than data point from common database with fixed size. Even outlier have an abnormal behaviour, often contain important information.
The task of outlier detection is to find small groups of data objects that are exceptional when compared with rest large amount of data. Detection of such outliers is important for many applications such as fraud detection, intrusion detection, and network monitoring application. Most existing methods are designed for numeric data. They will encounter problems with real-life applications that contain categorical data.
LSA method detects outlier, with a local-search heuristic based algorithm and entropy formulation for the calculation.

Keywords : data mining, outlier, outlier detection, LSA, entropy

1. Pendahuluan

1.1 Latar belakang

Perkembangan data yang pesat tidak lepas dari perkembangan tekhnologi informasi yang memungkinkan data dalam jumlah besar terakumulasi, ledakan data hampir terjadi di setiap penjuru dunia baik industri, instansi dan internet. Dengan kondisi seperti ini terdapat banyak tuntutan untuk menemukan informasi berguna yang tenggelam dalam tumpukan data dari berbagai sumber. Data dengan jumlah yang begitu besar ini akan sangat menyulitkan apabila kita ingin menganalisa apakah terdapat suatu kesalahan dalam data tersebut. Data yang mempunyai sifat dan karakteristik yang berbeda dari data pada umumnya dan mempunyai kemunculan kejadian relatif sedikit dikatakan sebagai *outlier*.

Outlier detection dapat didefinisikan sebagai pencarian terhadap sebagian kecil dari data, yang memiliki sifat yang berbeda jika dibandingkan dengan data keseluruhan. *Outlier* sendiri dapat didefinisikan sebagai sebuah titik data pada suatu basis data dimana sangat berbeda dibandingkan dengan titik data pada basis data pada umumnya. *Outlier* seringkali mempunyai informasi yang sangat berguna karena memiliki karakteristik yang tidak normal, sehingga dapat menghidupkan kecurigaan misalnya pada aplikasi kecurangan kartu kredit, *network intrusion detection*, aplikasi keuangan dan lain lain.

Dalam data mining terdapat banyak metoda dalam pencarian *outlier* seperti *clustering* yang mendefinisikan sebuah *outlier* tidak terdapat dalam *cluster* tersebut, dengan kata lain, *clustering* secara implisit mendefinisikan *outlier* sebagai *noise* dari suatu cluster tertentu. Teknik lainnya metode statistik dengan medefinisikan sebuah *outlier* berada diluar sekumpulan data yang ada. Metode *distance-based* mendefinisikan sebuah *outlier* berada jauh dari pusat data. Metode *density-based* mendefinisikan sebuah *outlier* merupakan sekumpulan titik data dengan kepadatan yang sangat rendah[15].

Permasalahannya adalah metoda-metoda yang ada seperti metode *clustering*, *m*etode *distance-based* dan metode *density-based* tidak mengkhususkan pencarian outlier pada data yang yang bersifat *categorical*, malah metoda tersebut kebanyakan untuk menangani data yang bersifat *numeric*, padahal dalam aplikasi didunia nyata terdapat data yang bersifat *categorical* bukan hanya *numeric*. Maka dengan metoda *LSA* algoritma akan memecahkan problem yang ada pada aplikasi di dunia nyata yang banyak mengandung data yang bersifat *categorical*[8].

1.2 Perumusan masalah

Masalah pokok yang akan diteliti adalah :

1. Bagaimana cara menguji algoritma *LSA* pada pendeteksian *outlier* pada *categorical* data.
2. Bagaimana analisa nilai perhitungan nilai *entropy* untuk mendeteksi *outlier*.
3. Bagaimana implementasi algoritma *LSA* pada aplikasi *outlier detection*.
4. Bagaimana analisa akurasi algoritma *LSA* dalam menentukan *outlier*.

Dalam penyusunan tugas akhir ini permasalahan dibatasi dalam beberapa hal, yaitu :

1. Algoritma yang digunakan adalah *Local Search Algorithm* (LSA).
2. Tidak menangani *preprosesing* data.
3. Tipe data berupa *categorical.*
4. Dataset bersifat *supervised.*
5. Input data yang akan dilakukan deteksi *outlier* adalah file *.arff.*
6. Pengujian dilakukan pada dataset dengan distribusi kelas yang *imbalanced* dan telah diketahui jumlah outliernya.
7. Analisis performansi meliputi akurasi obyek *outlier* yang dihasilkan, dan waktu eksekusi terhadap peningkatan jumlah data dan penambahan jumlah inputan k-outlier.

1.3 Tujuan

Tujuan dari penelitian tugas akhir ini adalah sebagai berikut :

1. Implementasi Algoritma *LSA* untuk mendeteksi *outlier* pada data *categorical*.
2. Membangun perangkat lunak deteksi *outlier* dengan menerapkan metode *LSA.*
3. Melakukan analisa terhadap perangkat lunak untuk menguji akurasi dan waktu deteksi terhadap penambahan jumlah data dan nilai inputan *k-outlier*.

1.4 Metodologi penyelesaian masalah

Metodologi yang akan digunakan dalam merealisasikan tujuan dan pemecahan masalah di atas adalah dengan menggunakan langkah-langkah berikut.

1. Studi pustaka
   Pada tahap ini dipelajari metoda LSA dalam pendektesian *outlier* pada *categorical* data, dan metoda – metoda lainnya seperti metode *clustering, m*etode *distance-based* dan metode *density-based*.
2. Analisis dan Desain
   Pada tahap ini dilakukan analisis pemecahan dari permasalahan pada data *categorical* yang akan dideteksi *outlier*-nya dengan menggunakan dasar teori yang telah dipelajari pada tahap sebelumnya.
3. Implementasi
   Hasil yang telah dilakukan pada tahap perancangan dapat diimplementasikan pada bahasa pemrograman.
4. Pengujian dan Evaluasi
   Menganalisis dengan mengukur akurasi pendeteksian *outlier* hasil implementasi metoda *LSA.*

5. Penutup

5.1 Kesimpulan

1. Perangkat lunak yang dibangun dengan algoritma *LSA* dapat mendeteksi *outlier* pada *categoric*al data.
2. Semakin *imbalanced* distribusi dataset, persentase akurasi semakin baik
3. Semakin bertambahnya nilai inputan k–*outlier*, persentase akurasi semakin baik, pada inputan k-outlier < 50% jumlah dataset dan juga berimbas pada, semakin lama waktu deteksi *outlier*.
4. Semakin besar jumlah data, semakin lama waktu deteksi *outlier*.
5. Perhitungan *entropy* dapat menjadi tolak ukur dalam mendeteksi *outlier* pada *categorical* data.
6. Nilai terbaik akurasi dari algoritma *LSA* mencapai 100% pada data dengan % rare class 4.1% dan 5.6% dengan inputan k-*outlier* sebanyak jumlah *outlier* yang terdapat pada dataset.
7. Distribusi *class* yang terurut dapat menghasilkan persentase akurasi yang lebih baik dan meminimalisasi jumlah inputan k-*outlier* optimal.

5.2 Saran

1. Diperlukan pengujian untuk menganalisa apakah algoritma *LSA* dapat digunakan untuk mendeteksi *outlier* pada data numerik.
2. Diperlukan penanganan khusus untuk mengatasi waktu deteksi yang lama pada data dengan jumlah yang besar.

Referensi

[1]	Aggarwal, C., Yu, P. S., Park, 2001, *Outlier detection for high dimensional data*, SIGMOD'01.
[2]	Barnett, V., Lewis, T., 1994, Outliers in Statistical Data, John Wiley ang Sons, New York .
[3]	Edwin M. Knorr, Raymond T. Ng, *Algorithms for Mining Distance-Based Outliers in Large Datasets*, In Proc. 24th Int. Conf. Very Large Data Bases, VLDB, 1998.
[4]	Fayyad, Usama. "Advances in Knowledge Discovery and Data Mining". MIT Press. 1996
[5]	Han, J., Kamber, M., 2001, *Data Mining: Concepts and Techniques*, USA: Morgan Kaufmann, Academic Press.
[6]	Hawkin, D., 1980, *Identification Of Outliers*, Chapman and Hall, Reading, London..
[7]	He, Z., Xu, X., J. Huang, J.Z., Deng. S., 2004, A Frequent Pattern Discovery Based Method for Outlier Detection. WAIM'04.
[8]	He, Z., Xu, X., J. Huang, J.Z., Deng. S., *An Optimization Model For Outlier Detection In Categorical Data*. WAIM
[9]	Lozano. Elio, Acuña. Edgar, *Parallel Algorithms for distance-based and density-based outliers*, University of Puerto Rico Mathematics Department.
[10]	Markus M. Breunig, Hans-Peter Kriegel, Raymod T. Ng, Jorg Sander, 2000, *LOF : Identifying Density-Based Local outliers*. In ACM SIGMOD International Conference on Management of Data .
[11]	G. J. Williams, R. A. baster, H. He, S. Harkins, L. Gu. A Comparative Study of RNN for Outlier Detection in Data Mining. ICDM'02. pp.2002
[12]	Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, dan Alesandro Zanasi, Discovering Data Mining: From Concept to Implementation, Prentice Hall, New Jersey, USA, 1998.
[13]	Shannon, C.E, 1984, *A Mathematical Theory of Communication*, Bell System Technical Journal.
[14]	Fayyad, Usama. "Advances in Knowledge Discovery and Data Mining". MIT Press. 1996
[15]	Dedy Handriyadi. 2006. *ANALISA PERBANDINGAN CLUSTERING-BASED, DISTANCE-BASED DAN DENSITY-BASED DALAM MENDETEKSI OUTLIER*. STT Telkom. Bandung.