 ISBN : 978-602-97094-4-5

Deteksi *Outlier* Menggunakan Algoritma *Naive Nested Loop* (Studi Kasus : Data Akademik Mahasiswa Program Studi PS Universitas XYZ)

Setyo Resmi Probowati[1], Paulina H. Prima Rosa[2]
[1,2]Jurusan Teknik Informatika, Universitas Sanata Dharma, Yogyakarta
setyoresmi@gmail.com, rosa @usd.ac.id

***Abstrak* — Dalam makalah ini diuraikan penelitian yang dimaksudkan untuk membangun sistem yang dapat melakukan deteksi *outlier* terhadap data numerik dengan menerapkan algoritma *Naïve Nested Loop*. Data yang dipergunakan dalam penelitian adalah data akademik mahasiswa program studi PS Universitas XYZ, Yogyakarta tahun angkatan 2007 dan 2008. Data tersebut terdiri dari data numerik nilai hasil seleksi masuk mahasiswa yang diterima melalui jalur tes tertulis maupun jalur prestasi dan nilai indeks prestasi dari semester satu sampai empat. Dari hasil pengujian *review* dan validitas oleh pengguna dapat disimpulkan bahwa sistem dapat menghasilkan data yang dinyatakan sebagai *outlier*. Sedangkan berdasar hasil pengujian efek perubahan nilai atribut penambangan data disimpulkan bahwa penentuan nilai parameter M dan dmin pada algoritma *Naïve Nested Loop* berpengaruh terhadap jumlah *outlier* yang dihasilkan.**

Kata kunci* — penambangan data, deteksi *outlier, Naïve Nested Loop.

I. PENDAHULUAN

A. Latar Belakang

Penambangan data merupakan ekstraksi pola terhadap data yang menarik dalam jumlah yang besar. Pola tersebut dikatakan menarik apabila tidak diketahui sebelumnya dan berguna bagi perkembangan ilmu pengetahuan. Sementara itu, data tersebut dapat diolah dengan berbagai teknik penambangan data seperti asosiasi, klasifikasi, *clustering* dan deteksi *outlier*.

Outlier merupakan sebuah data yang berbeda dibandingkan dengan sifat umum yang dimiliki data lain pada suatu kumpulan data. *Outlier* juga memiliki fungsi untuk mendeteksi perilaku yang tidak normal seperti deteksi penyalahgunaan kartu kredit, deteksi adanya penyusupan pada jaringan komunikasi, analisis medis, segmentasi data pelanggan yang berkaitan dengan pemasaran barang. Dari hal tersebut, penulis tertarik melakukan penelitian mengenai deteksi *outlier* menggunakan algoritma *Naïve Nested Loop* terhadap data akademik mahasiswa program studi PS Universitas XYZ tahun angkatan 2007 dan 2008. Dari hasil penelitian diharapkan dapat dihasilkan informasi yang menarik mengenai data mahasiswa yang berbeda dari data lainnya.

.

B. Metodologi

Dalam penelitian ini dilakukan tahap-tahap sebagai berikut:

1. Mempelajari langkah-langkah yang berkaitan dengan deteksi *outlier* menggunakan algoritma *Naïve Nested Loop*.
2. Mengumpulkan dan mengelola data mahasiswa program studi PS, Universitas XYZ, angkatan 2007 dan 2008. Data tersebut didapat dari hasil penelitian yang dilakukan oleh Rosa dkk [1]. Kemudian, data tersebut dikelompokkan menjadi tiga *dataset* berdasarkan tes penerimaan masuk mahasiswa. Tiga *dataset* tersebut dapat dikelompokkan sebagai berikut:
   (1) DATASET1: Mahasiswa program studi PS angkatan 2007 dan 2008 yang diterima melalui jalur tes tertulis.
   (2) DATASET2: Mahasiswa program studi PS angkatan 2007 dan 2008 yang diterima melalui jalur prestasi.
   (3) DATASET3: Mahasiswa program studi PS angkatan 2007 dan 2008 yang diterima melalui jalur tes tertulis dan prestasi.
3. Penerapan implementasi sistem pendeteksi *outlier* menggunakan algoritma *Naïve Nested Loop*. Sistem tersebut dibangun dan dikembangkan dengan bahasa pemrograman berbasis Java.
4. Melakukan dua jenis pengujian terhadap dataset, yaitu:
   (1) *Review* & validitas oleh pengguna, dalam hal ini dilakukan oleh Ketua Program Studi PS, yang digunakan untuk melihat kesesuaian antara hasil yang dikeluarkan oleh sistem dengan hasil yang diharapkan oleh pengguna.
   (2) Uji efek perubahan nilai atribut penambangan data adalah tahap untuk melihat pengaruh penentuan nilai parameter M dan dmin terhadap jumlah *outlier* yang dihasilkan.

II. LANDASAN TEORI

A. Penambangan Data

Definisi umum dari penambangan data adalah serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data [2]. Penambangan data mengekstraksi pola yang menarik dari data dalam jumlah besar. Suatu pola dikatakan menarik apabila pola tersebut tidak *sepele*, implisit, tidak diketahui sebelumnya, dan berguna.

Berikut fungsionalitas dan tipe pola yang dapat ditemukan dengan penambangan data [3]:

1. Deskripsi konsep/ kelas
   Hal ini bermanfaat untuk mendeskripsikan masing-masing kelas atau konsep tersebut dengan deskripsi kelas atau konsep.
2. Analisis Asosiasi (Korelasi dan kausalitas)
   Analisis asosiasi adalah pencarian aturan-aturan asosiasi yang menunjukkan kondisi-kondisi nilai atribut yang sering terjadi bersama-sama dalam sekumpulan data.
3. Klasifikasi dan Prediksi
   Klasifikasi adalah proses menemukan model atau fungsi yang menjelaskan dan membedakan kelas-kelas atau konsep, dengan tujuan agar model yang diperoleh dapat digunakan untuk memprediksi kelas atau objek yang memiliki label kelas yang tidak diketahui.
4. Analisis Klaster
   Tidak seperti klasifikasi dan prediksi, yang menganalisis objek data yang diberi label kelas, *clustering* menganalisis objek data di mana label kelas tidak diketahui.
5. Analisis *Outlier*
   Database dapat mengandung objek data yang tidak sesuai dengan sifat umum atau model data. Objek data tersebut adalah *outlier*. *Outlier* merupakan objek data yang tidak mengikuti perilaku umum dari data.
6. Analisis Evolusi
   Analisis evolusi data menjelaskan dan memodelkan tren dari objek yang memiliki perilaku yang berubah setiap waktu. Teknik ini dapat meliputi karakteristik, diskriminasi, asosiasi, klasifikasi, atau *clustering* dari data yang berkaitan dengan waktu.

B. Outlier

Outlier merupakan kumpulan data yang dianggap memiliki sifat yang berbeda, tidak konsisten dibandingkan dengan kebanyakan data lainnya [3]. *Outlier* sering dianggap sebagai *noise* , namun untuk kasus-kasus tertentu justru informasi yang tidak konsisten tersebut bisa dikatakan lebih menarik dan bermanfaat. Meskipun dapat dikatakan bahwa *outlier* merupakan data yang cukup berbeda dari data lain, masih sedikit persetujuan yang menyatakan seperti apakah ketentuan *outlier* yang bermakna [4].

Deteksi *outlier* merupakan salah satu bidang penelitian yang penting dalam topik penambangan data. Penelitian ini bermanfaat untuk mendeteksi perilaku yang tidak normal seperti deteksi penyalahgunaan kartu kredit, deteksi adanya penyusupan pada jaringan komunikasi, analisis medis, segmentasi data pelanggan yang berkaitan dengan pemasaran barang [3]. Banyak metode telah dikembangkan untuk menyelesaikan masalah ini, namun kebanyakan hanya fokus pada data dengan atribut yang seragam, yaitu data numerik atau data kategorikal saja.

Outlier dapat disebabkan karena data berasal dari sumber yang berbeda, variasi alami dari data itu sendiri, dan kesalahan saat pengukuran atau eksekusi data [3]. Adanya data *outlier* ini akan membuat analisis terhadap serangkaian data menjadi bias, atau tidak mencerminkan fenomena yang sebenarnya. Istilah *outlier* juga sering dikaitkan dengan nilai esktrem, baik ekstrem besar maupun ekstrem kecil.

C. Algoritma Naive Nested Loop

Prinsip kerja algoritma *Naïve Nested Loop* adalah mendeteksi *outlier* pada sekumpulan data lalu mencari tetangga untuk masing-masing objek dalam radius *dmin* disekitaran objek tersebut. M adalah jumlah maksimum objek dalam ketetanggaan *dmin* dari sebuah *outlier* dan *dmin* adalah radius atau jarak maksimum ketetanggaan antar objek o.

Dalam penelitian yang dilakukan oleh Knoor [4] nilai M juga dinyatakan sebagai n(1-p), di mana n merupakan jumlah data, *p* atau disebut juga *pct* merupakan jumlah minimum objek yang terletak lebih jauh dari jarak o ke *dmin*.

Berikut merupakan cara kerja algoritma *Naïve Nested Loop* [5]:

for j = 1 to n do

- *set countj = 0 ;*
- *for k=1 to n do if (dist(j,k)<D) then countj++ ;*
- *if countj <= |n(1-p)| then output j as an outlier*, di mana dist merupakan jarak antara objek j dengan k, dan D yang bernilai sama dengan dmin adalah radius atau jarak maksimum ketetanggaan antar objek o. Sedangkan nilai n(1-p) sama dengan M.

III. PERANCANGAN DAN IMPLEMENTASI

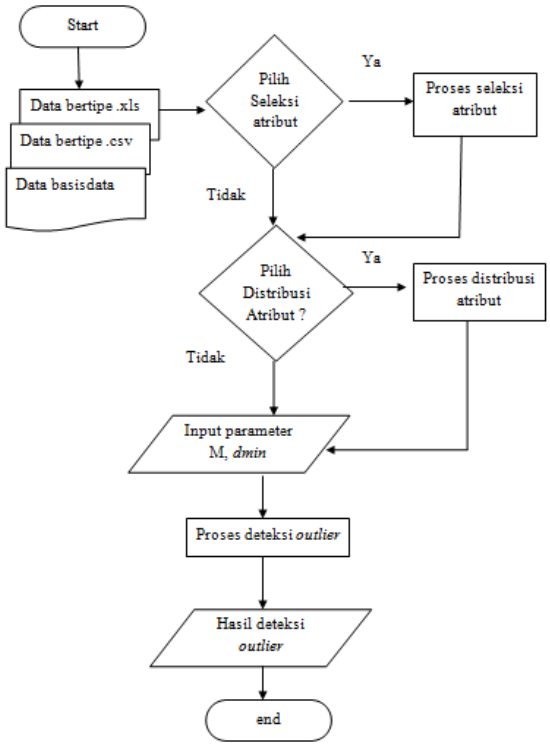
Untuk membantu dalam melakukan pengujian, maka dibutuhkan sistem yang mampu mendeteksi *outlier* menggunakan algoritma *Naïve Nested Loop* dengan data mahasiswa Program Studi PS Universitas XYZ. Sistem tersebut dibangun dan dikembangkan dengan bahasa pemrograman berbasis Java.

Terdapat tiga fungsi utama yang dapat dijalankan oleh pengguna yaitu fungsi *input* data, fungsi pendeteksian *outlier*, dan fungsi simpan hasil deteksi *outlier*. Ketiga fungsi tersebut saling berkaitan sehingga dalam menjalankan fungsi ini pengguna harus melakukannya secara berurutan. Gambar 1 menjelaskan proses umum sistem pendeteksi *outlier* yang dilakukan secara terurut untuk dapat menjalankan ketiga fungsi tersebut. Bentuk *file* data yang digunakan sebagai masukkan pada sistem ini adalah *.xls, .csv* dan tabel dari basis data (Oracle dan MySql). Sementara itu, hasil keluaran yang diharapkan adalah hasil outlier beserta nilai atribut yang dimiliki.

Masukan dalam bentuk *file* (*.xls, .csv* dan tabel dari basis data) tersebut ditampilkan ke dalam tabel *view*, sehingga pengguna dapat melakukan fungsi seleksi atribut atau melihat grafik distribusi per atribut.

Pengguna harus memasukan nilai parameter M dan dmin terlebih dahulu sebelum mendapatkan hasil *outlier* yang

diinginkan. Hasil *outlier* tersebut dapat simpan dalam bentuk .doc atau .txt.



Gambar 1.Proses umum sistem pendeteksi outlier menggunakan algoritma *Naïve Nested Loop* [6].

IV. ANALISA HASIL

Pengujian pada penelitian ini dilakukan menggunakan dua jenis pengujian. Uji pertama yaitu *review* & validasi oleh pengguna yang digunakan untuk mengevaluasi apakah *outlier* yang ditemukan oleh sistem sungguh-sungguh merupakan *outlier* dari sudut pandang pengguna (Kaprodi PS). Sementara uji kedua adalah uji efek perubahan nilai atribut penambangan data untuk melihat pengaruh penentuan nilai parameter M dan dmin terhadap jumlah *outlier* yang dihasilkan,

Untuk keperluan uji pertama, berikut adalah atribut-atribut yang digunakan pada setiap *dataset*.

1. DATASET1 berisi 72 data mahasiswa dan menggunakan atribut ips1, ips2, ips3, ips4, nil1, nil2, nil3, nil4, dan nil5.
2. DATASET1 berisi 54 data mahasiswa dan menggunakan atribut ips1, ips2, ips3, ips4, dan final.
3. DATASET1 berisi 126 data mahasiswa dan menggunakan atribut ips1, ips2, ips3, ips4 dan nilai final.

Tabel 1a-1c menjelaskan hasil *outlier* untuk setiap *dataset* data mahasiswa.

Tabel 1a. Hasil deteksi *outlier* dari DATASET1 dengan nilai M = 5 dan dmin = 2.

Semester	No.Urut Mahasiswa	ips1	ips2	ips3	ips4	nil1	nil2	nil3	nil4	nil5
1	2	1.72	-	-	-	1.20	0.80	3.20	1.20	0.40
	8	1.44	-	-	-	4.00	2.00	3.60	2.40	2.80
2	2	-	1.65	-	-	1.20	0.80	3.20	1.20	0.40
	52	-	1.28	-	-	2.40	3.20	3.60	2.80	3.20
	54	-	1.04	-	-	3.20	2.00	3.60	3.20	1.60
3	2	-	-	1.53	-	1.20	0.80	3.20	1.20	0.40
	30	-	-	0.59	-	2.40	2.40	2.80	1.20	3.60
	54	-	-	0.19	-	3.20	2.00	3.60	3.20	1.60
4	2	-	-	-	1.68	1.20	0.80	3.20	1.20	0.40
	16	-	-	-	1.07	2.40	1.60	2.80	3.60	2.80
	27	-	-	-	0.05	2.80	2.40	2.40	2.00	2.00
	47	-	-	-	0.00	2.80	2.00	1.20	1.60	1.60
	48	-	-	-	0.00	2.80	1.60	2.00	2.40	1.20
	52	-	-	-	0.82	2.40	3.20	3.60	2.80	3.20

Tabel 1b. Hasil deteksi *outlier* dari DATASET2 dengan nilai M= 5 dan dmin = 1.

Semester	No.Urut Mahasiswa	ips1	ips2	ips3	ips4	Final
1	76	0.85	-	-	-	2.85
2	71	-	0.32	-	-	2.72
	124	-	0.44	-	-	2.90
3	71	-	-	0.69	-	2.72
4	88	-	-	-	0.94	2.89

Tabel 1c. Hasil deteksi *outlier* dari DATASET3 dengan nilai M= 5 dan dmin = 1.

Semester	No. Urut Mahasiswa	ips1	ips2	ips3	ips4	Final	Jalur Tes
1	2	1.72	-	-	-	1.12	Tes
	76	0.85	-	-	-	2.85	Prestasi
2	71	-	0.32	-	-	2.72	Prestasi
	2	-	1.65	-	-	1.12	Tes
	124	-	0.44	-	-	2.90	Prestasi
3	2	-	-	1.53	-	1.12	Tes
	71	-	-	0.69	-	2.72	Prestasi
	30	-	-	0.59	-	2.56	Tes
	54	-	-	0.19	-	2.56	Tes
4	2	-	-	-	1.68	1.12	Tes
	47	-	-	-	0.00	2.04	Tes
	48	-	-	-	0.00	2.00	Tes

Seluruh hasil *outlier* telah dikonfirmasi oleh Kaprodi dan dinyatakan bahwa data-data *outlier* yang menjadi *output* sistem adalah juga *outlier* di mata Kaprodi. Dengan demikian, dapat disimpulkan bahwa sistem telah berhasil melakukan deteksi *outlier* terhadap ketiga dataset tersebut.

Pada uji kedua, terhadap ketiga dataset dilakukan pengujian untuk melihat efek perubahan nilai parameter M dan dmin terhadap jumlah *outlier* yang dihasilkan. Tabel 2a-2c mendeskripsikan jumlah *outlier* yang ditemukan pada ketiga dataset dengan berbagai macam nilai M dan D.

Tabel 2a. Variasi Jumlah *outlier* DATASET1 dengan nilai M dan dmin yang berubah-ubah.

	M = 1	M = 2	M = 3	M = 4	M = 5
dmin = 1	27	32	36	42	46
dmin = 2	1	2	2	2	2

dmin = 3	0	0	0	0	0
dmin = 4	0	0	0	0	0

Tabel 2b. . Variasi Jumlah *outlier* DATASET2 dengan nilai M dan dmin yang berubah-ubah

	M = 1	M = 2	M = 3	M = 4	M = 5
dmin = 1	0	1	1	1	1
dmin = 2	0	0	0	0	0
dmin = 3	0	0	0	0	0
dmin = 4	0	0	0	0	0

Tabel 2c. . Variasi Jumlah *outlier* DATASET3 dengan nilai M dan dmin yang berubah-ubah.

	M = 1	M = 2	M = 3	M = 4	M = 5
dmin = 1	0	1	1	1	1
dmin = 2	0	0	0	0	0
dmin = 3	0	0	0	0	0
dmin = 4	0	0	0	0	0

Berdasarkan ketiga hasil percobaan di atas, nampak bahwa nilai parameter M dan dmin pada algoritma *Naïve Nested Loop* berpengaruh terhadap jumlah hasil *outlier* yang dihasilkan. Jika nilai M tetap dan nilai dmin bertambah maka jumlah *outlier* bisa tetap atau berkurang. Sementara itu jika dmin tetap dan nilai M bertambah maka jumlah *outlier* bisa tetap atau bertambah.

V. SIMPULAN

Berdasarkan hasil pengujian dapat disimpulkan bahwa algoritma *Naïve Nested Loop* dapat diimplementasikan untuk menemukan data *outlier* dari sekumpulan dataset numerik, berupa data nilai yang berbeda dari data lainnya sehingga data tersebut dinyatakan sebagai *outlier*. *Outlier* yang dihasilkan telah dikonfirmasi oleh pengguna sebagai *outlier* pula. Selain itu, nilai parameter pada algoritma *Naïve Nested Loop* yaitu nilai M dan dmin mempengaruhi jumlah *outlier* yang dihasilkan.

Penelitian ini dimungkinkan untuk dikembangkan lebih lanjut, salah satunya adalah untuk mendeteksi *outlier* pada data yang bersifat kategorikal maupun data campuran, bukan hanya data numerik.

DAFTAR PUSTAKA

[1] P.H. P. Rosa, R. Gunawan, S.H. Wijono, "The Development of Academic Data Warehouse as a Basis for Decision Making : A Case Study at XYZ University", *Proceeding of International Conference on Enterprise Information System and Application, Yogyakarta: Universitas Islam Indonesia*, 2013.

[2] I. Pramudiono, *Pengantar Data Mining : Menambang Permata Pengetahuan di Gunung Data.* http://www.ilmukomputer.org/wp-content/uploads/2006/08/iko-datamining.zip. Diakses pada tanggal 15 November 2012 jam 21.32.

[3] J. Han, & M. Kamber *Data Mining Concepts and Techniques Second Edition*. San Fransisco : Elsovier. 2006

[4] E. M. Knorr, and Raymond T. Ng, "Algorithms for Mining Distance-Based Outliers in Large Datasets," *Proceedings of the 24rd International Conference on Very Large Data Bases*, hal. 392-403, 1998.

[5] J. Pei, *CMTP 741 Foundation of DataMining–OutlierDetection.pdf*,http://www.cs.sfu.ca/CourseCentral/741/jpei/slides/. Diakses pada tanggal 22 Agustus 2013 jam 15.00.

[6] S.P. Probowati, *Deteksi Outlier Menggunakan Algoritma Naive Nested Loop*. Skripsi tidak diterbitkan. Yogyakarta: Universitas Sanata Dharma, 2013.