*Seminar I*

**PERBANDINGAN METODE-METODE ONE CLASS CLASSIFICATION   
PADA KELAS TIDAK SEIMBANG**



***Oleh :***

**BAGAS PRASETYO**

**H13116308**

**Pembimbing Utama :   
Pembimbing Pertama :   
Penguji :**

**PROGRAM STUDI ILMU KOMPUTER**

**DEPARTEMEN MATEMATIKA**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS HASANUDDIN**

**2019**

# **BAB I PENDAHULUAN**

## **Latar Belakang**

Data tidak seimbang merupakan suatu keadaan dimana distribusi kelas data tidak seimbang, jumlah kelas data (instance) yang satu lebih sedikit atau lebih banyak dibanding dengan jumlah kelas data lainnya. Kelompok kelas data yang lebih sedikit dikenal dengan kelompok minoritas (minority), kelompok kelas data yang lainnya disebut dengan kelompok mayoritas (mayority). Kondisi tersebut akan berpengaruh terhadap klasifikasi data yang akan dilakukan untuk menentukan kelas suatu data. Jika kondisi data tersebut tidak seimbang(imbalance) maka kecenderungan kelas data tidak stabil karena data akan lebih condong ke bagian data yang memiliki komposisi data lebih besar (Majority class).

Proses klasifikasi dengan berbagai algoritma machine learning, yaitu algoritma yang mempelajari pengelompokan data berdasarkan pengolahan data yang telah ada sebelumnya, bertujuan untuk mendapatkan target kelas yang akurat. Tapi hampir semua algoritma klasifikasi seperti Naive Bayes, Decision Tree, K-Nearest Neighbor dan yang lainnya menunjukkan performa yang sangat buruk ketika bekerja pada data dengan kelas yang sangat tidak seimbang. Jika digunakan pada data tidak seimbang algoritma klasifikasi akan menghasilkan akurasi yang sangat tinggi untuk kelas mayoritas daripada kelas minoritas dan spesifisitas sangat rendah.

Masalah klasifikasi satu kelas menjadi sangat penting dalam penelitian pembelajaran mesin baru-baru ini. Satu perbedaan mendasar antara klasifikasi satu kelas (One Class Classification) dan klasifikasi konvensional adalah bahwa dalam pembelajaran satu kelas, diasumsikan bahwa hanya informasi kelas target yang tersedia. Dengan kata lain, dalam proses pelatihan classifier, instance dari kelas target digunakan dan tidak ada informasi tentang rekannya. Batas antara kedua kelas harus diperkirakan dari data dari satu-satunya objek yang tersedia. Dengan demikian, tugasnya adalah untuk menentukan batas di sekitar kelas target, sehingga ia melingkari sebanyak mungkin contoh target dan meminimalkan kemungkinan menerima outlier.

Klasifikasi satu kelas cocok pada dataset pelatihan yang hanya memiliki contoh dari kelas normal. Setelah disiapkan, model ini digunakan untuk mengklasifikasikan contoh baru sebagai normal atau tidak normal, yaitu outlier atau anomali. Teknik klasifikasi satu kelas dapat digunakan untuk masalah klasifikasi tidak seimbang biner (dua kelas) di mana kasus negatif (kelas 0) diambil sebagai " normal " dan kasus positif (kelas 1) diambil sebagai pencilan atau anomali.

Menerapkan klasifikasi satu kelas untuk data yang tidak seimbang adalah arah penelitian yang jarang dilakukan, meskipun beberapa pekerjaan telah dilakukan. Karena karakteristik ketidakseimbangan data, klasifikasi konvensional cenderung memiliki bias terhadap kelas mayoritas. Keuntungan dari klasifikasi satu kelas adalah bahwa dengan membuang mayoritas yang mengganggu, "ruang" tempat data minoritas berada dapat ditentukan dengan lebih baik. Meskipun tidak dirancang untuk jenis masalah ini, algoritma klasifikasi satu kelas dapat efektif untuk dataset yang tidak seimbang di mana tidak ada atau sangat sedikit contoh kelas minoritas.

Berdasarkan uraian di atas, maka peneliti akan perbandingan algoritma-algoritma klasifikasi satu kelas (one class classification) untuk membuat model klasifikasi pada data tidak seimbang dengan judul “Perbandingan algoritma-algoritma One Class Classification pada data tidak seimbang”

## **Rumusan Masalah**

Adapun rumusan masalah dalam penelitian ini adalah :

1. Bagaimana implementasi algoritma-algoritma one class classification pada data tidak seimbang?
2. Bagaimana performa algoritma-algoritma one class classification dalam membuat model pada data tidak seimbang?
3. Algoritma one class classifiation manakah yang memiliki performa terbaik pada dataset tidak seimbang?

## **Batasan Masalah**

Batasan masalah pada penelitian ini adalah :

1. Dataset yang digunakan adalah dataset tak seimbang.
2. Dataset hanya memiliki dua kelas (biner).
3. Algoritma One Class Classification yang digunakan adalah One Class SVM, Isolation Forest, and Local Outlier Factor.

## **Tujuan Penelitian**

Berdasarkan rumusan masalah, maka tujuan dari penelitian ini adalah:

1. Mengetahui implementasi algoritma-algoritma one class classification pada dataset tidak seimbang.
2. Mengetahui performa algoritma-algoritma one class classification yang dihasilkan dari dataset tidak seimbang.
3. Mengetahui algoritma one class classification yang memiliki performa terbaik pada dataset tidak seimbang.

## **Manfaat Penelitian**

Hasil penelitian ini diharapkan dapat bermanfaat:

1. Sebagai rujukan untuk mengatasi dataset tak seimbang yang sangat sering dijumpai.
2. Menjadi sumber informasi mengenai performa algoritma-algoritma one class classification pada dataset tidak seimbang.
3. Menjadi sumber informasi mengenai performa dari algoritma-algoritma one class classification pada dataset tidak seimbang.

# **BAB II TINJAUAN PUSTAKA**

## ***Machine Learning***

*Machine learning* adalah memprogram komputer untuk mengoptimalkan suatu ukuran kinerja menggunakan sampel data atau berdasarkan pengalaman (Ethem, 2009). *Machine learning* menggunakan suatu algoritma untuk menganalisis data.

Dalam pembelajaran yang terawasi atau *supervised learning*, pengklasifikasi akan diberikan suatu input tertentu dan menghubungkannya dengan suatu output. Kasus di mana tujuannya adalah mengklasifkasikan input data ke suatu kategori diskrit tertentu disebut *klasifikasi*, dan kasus di mana outputnya adalah suatu variabel kontinu disebut *regresi*. Dalam pembelajaran tanpa pengawasan atau *unsupervised learning*, pengklasifikasi diberi input dan dibiarkan sendiri untuk menemukan pola pada data tersebut. Kasus unsupervised learning di mana tujuannya adalah mengelompokkan observasi-observasi yang mirip disebut *clustering*, jika menentukan distribusi data pada input disebut *estimasi kepadatan*. Dalam *reinforcement learning*, sistem komputer menerima input secara terus menerus dan mencoba memilih keputusan-keputusan yang paling optimal berdasarkan kondisi lingkungannya.

Masing-masing jenis pembelajaran memiliki banyak algoritma yang telah dikembangkan dengan pendekatan yang berbeda-beda.

## **Imbalanced Class**

Dataset adalah kumpulan data yang berbentuk tabel, di mana setiap kolomnya merepresentasikan suatu ciri-ciri, atribut atau fitur. Setiap barisnya menyatakan observasi suatu individu, record atau sampel (Snijders, Matzat, & Reips, 2012). Suatu dataset biasanya memiliki satu kolom tambahan yang merepresentasikan kelas dari observasi tersebut, kolom ini disebut kolom kelas. Kolom kelas ini juga disebut sebagai variabel dependen terhadap variabel-variabel independen yang merupakan ciri-ciri (atribut) dari suatu observasi tertentu.

Dalam *machine learning* dikenal istilah dataset dengan class yang tak seimbang. Istilah ini berlaku ketika kelas dari dataset tersebut bersifat kategorik diskrit. Dataset dengan class yang tak seimbang (*imbalanced class*) adalah dataset yang frekuensi kejadian dari kelas tertentu sangat jauh berbeda dengan kelas yang lain. Contohnya seperti suatu dataset dengan jumlah pasien yang berkelas “diabetes” jumlahnya jauh lebih sedikit dibanding pasien yang “tidak diabetes”.

Masalah ketidakseimbangan ini akan memberi bias terhadap performa pengklasifikasi sebab jumlah sampel pada kelas tertentu tidak dapat memberi informasi yang cukup kepada pengklasifikasi berdasarkan ciri-ciri yang diberikan.

## ***Confusion Matrix***

Di dalam machine learning, mengukur kinerja atau performa dari suatu model adalah hal yang esensial. Model yang diperoleh dari pelatihan melalui data training perlu diuji melalui data testing. Kinerja diukur berdasarkan seberapa baik model tersebut memprediksi dengan benar data yang ada.

Pada klasifikasi biner, kelas positif yang berhasil diprediksi dengan benar disebut true positive, jika kelas positif tersebut diprediksi negatif (salah) disebut false negative. Kelas negatif yang berhasil diprediksi negatif (benar) disebut true negative, dan kelas negatif yang diprediksi positif disebut false positive. Jumlah dari kasus-kasus tersebut direpresentasikan dalam suatu tabel kontingensi yang disebut confusion matrix (Swets, 1988).

Tabel 1: Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | Kelas asli | | |
| Hasil prediksi |  | Positif | Negatif |
| Positif | TP | FP |
| Negatif | FN | TN |

Akurasi adalah ukuran kinerja yang menunjukkan seberapa baik suatu pengklasifikasi dalam mengklasifikasikan seluruh data. Akurasi adalah rasio antara observasi yang diklasifikasikan secara benar dengan total observasi:

|  |  |  |
| --- | --- | --- |
|  |  | (2.1) |

Presisi adalah ukuran kinerja yang menunjukkan seberapa besar kebenaran suatu pengklasfikasi dari seluruh kelas positif yang diprediksi. Presisi adalah rasio antara jumlah kelas positif yang diklasifikasikan secara benar dengan jumlah observasi yang diklasifikasikan positif:

|  |  |
| --- | --- |
|  | (2.2) |

*Recall* atau sensitivitas adalah ukuran kinerja yang menunjukkan seberapa baik suatu pengklasifikasi dalam mengklasifikasikan kelas positif. *Recall* adalah rasio antara jumlah observasi positif yang diklasifikasikan secara benar dengan jumlah observasi positif asli:

|  |  |  |
| --- | --- | --- |
|  |  | (2.3) |

F1-Score adalah *harmonic mean* antara *precision* dan *recall*:

|  |  |  |
| --- | --- | --- |
|  |  | (2.4) |

## **One Class Classification**

### ***One Class Support Vector Machine***

Formulasi SVM reguler adalah memaksimalkan margin garis pemisah antara dua kelas data. Untuk SVM dengan pendekatan one-class classification, (Schölkopf, Williamson, Smola, Shawe-Taylor, & Platt) memodifikasi SVM menjadi pencarian hiperplane dengan jarak pemisah maksimal antara titik-titik data yang diberikan (data positif) dengan titik origin. Hasil training berupa fungsi biner yang memetakan apakah suatu titik uji memiliki probabilitas tinggi berada di daerah data training (positif) atau di luar daerah (negatif). Pada formulasi One-class SVM ini ada tambahan parameter yaitu *v* (nu) yang bernilai antara 0 dan 1. Nilai *v* ini menentukan batas maksimal fraksi outlier pada dataset training, sekaligus batas minimal fraksi jumlah support vector terhadap jumlah data training.

Tingkat keakuratan metode One-Class SVM dalam memprediksi bergantung pada optimasi nilai parameter yang digunakan dalam suatu penelitian, parameter tersebut berupa gamma dan nu. Parameter gamma bernilai harus lebih besar dari 0, sedangkan parameter nu bernilai di bawah 1. Untuk mendapatkan nilai akurasi yang terbaik, optimasi parameter diperlukakn. Langkah-langkah yang dilakukan yaitu Training One-Class SVM dan Testing One-Class SVM. Training One-Class SVM menggunakan dataset training dengan berbagai nilai gamma (γ) dan nu (*v*).

### ***Isolation forest***

### ***Local outlier factor***

Ide utama dari LOF adalah membandingkan kepadatan lokal lingkungan sebuah objek dengan kepadatan lokal tetangganya, dalam algoritma LOF sebuah objek dikatakan sebagai outlier apabila memiliki nilai LOF yang tinggi atau menjauhi 1, sedangkan objek yang memiliki nilai LOF rendah atau mendekati 1 maka objek tersebut tidak dapat dikategorikan sebagai outlier. Nilai LOF yang tinggi mengindikasikan bahwa objek tersebut memiliki kepadatan yang rendah terhadap lingkungannya, sehingga berpotensi menjadi outlier. Langkah-langkah untuk menemukan outlier di deskripsikan sebagai berikut :

1. Menghitung *k-distance* dari setiap objek p

Tujuan dari perhitungan *k-distance* ini adalah untuk menentukan tetangga dari p, secara sederhana *k-distance* dari sebuah objek p adalah jarak maksimal dari p terhadap tetangga terdekatnya dan dinotasikan dengan *k-distance(p), k-distance* didefinisikan d(p,o) dimana antara p dan objek o ∈ D memiliki :

1. Untuk setidaknya *k* objek o’ ∈ D | {p} dan dinyatakan bahwa *d*(p,o’) ≤ *d*((p,o)
2. Untuk setiap *k*-1 objek o’ ∈ D | {p} dan dinyatakan bahwa *d*(p,o’) ≤ *d*((p,o)
3. Menghitung tetangga terdekat yang tidak lebih dari nilai *k-distance*

*k-deistance neighborhood* suatu objek p dinotasikan *Nk-distance(p)*, atau *Nk(p)* dimana berisi setiap objek dengan jarak lebih besar dari *k-distance(p).*

1. Menghitung *Reachability Distance* dari p, mencari jarak maksimum antar objek

*Reachability Distance* dari suatu objek p terhadap objek o adalah *distance(p, o)* atau *k-distance(o),* dengan membandingkan keduanya dan dicari nilai yang maksimum sehingga *reach-distk(p,o) = max{k-distance (o), distance(po)}.* Tujuan dari perhitungan ini adalah untuk memastikan bahwa semua benda berada dalam lingkungan yang homogen. Selain itu, nilai LOF akan stabil jika sebuah objek berada dalam lingkungan yang seragam sekalipun jika *MinPts(k)* berubah. Fluktuasi dari *reachability distance* dapat dikontrol dengan memilih nilai maksimum untuk k.

1. Menghitung *local reachability density* (kepadatan lokal) dari setiap objek

Secara intuitif, local reachbility density p adalah kebalikan dari rata-rata reachability distance based didalam minPts-tetangga terdekat dari objek p.

1. Menghitung LOF untuk setiap objek data

LOF adalah sebuah derajat yang menentukan apakah sebuat objek merupakan outlier atau tidak. dan merupakan rata-rata rasio kepadatan *reachability* lokal p dan tetangga p dalam satu jangkauan.

# **BAB III METODE PENELITIAN**

## **Waktu dan Tempat**

Penelitian ini dilaksanakan dari bulan Maret 2020 sampai dengan bulan April 2020. Lokasi penelitian dilakukan di Laboratorium Rekayasa Perangkat Lunak Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin.

## **Tahapan Penelitian**

Untuk menyelesaikan penelitian ini, peneliti akan melewati beberapa tahap penelitian, yaitu: Pra-penelitian, eksplorasi dan preprocessing data, model *tuning* dan *fitting* dan analisis hasil.

## **Deskripsi Data**

Data diambil dari Website resmi Kaggle (kaggle.com) yaitu:

1. Credit Card Fraud Dataset (Kaggle), yang terdiri dari 30 kolom attribut dengan 1 kolom kelas, 284.807 baris. 284.315 jumlah sampel kelas mayoritas dan 492 jumlah sampel kelas minoritas dengan imbalanced ratio sebesar 577:1. Dataset ini merupakan dataset terpopuler di Kaggle sebab jumlah data yang besar dengan imbalanced ratio yang sangat tinggi.

Dataset hanya memiliki atribut kontinu dengan label kelas biner yang sesuai dengan tema penelitian dan metode-metode yang digunakan.

## **Alur Penelitian**

# **Daftar Pustaka**

Azkiya, Zida Ziyan, Fatma Indriani, and Heru Kartika Chandra. "Deteksi Penyakit Dengue Hemorrhagic Fever dengan Pendekatan One Class Classification." *Journal of Information Systems Engineering and Business Intelligence* 3.2 (2017): 129-133.