J M S K
Jurnal Matematika, Statistika, & Komputasi
http://journal.unhas.ac.id/index.php/jmsk
Vol. 17, No. 1, 42-49, September, 2020
doi: 10.20956/jmsk.v%vi%i.10899

p-ISSN: 1858-1382
e-ISSN: 2614-8811

Comparison of Elliptic Envelope and Isolation Forest Algorithm on Imbalance Dataset

Komparasi Algoritma *Elliptic Envelope* dan *Isolation Forest* Pada Data Tidak Seimbang

Supri Amir[1*], Bagas Prasetyo[2]

Abstract

The problem of imbalanced data is important in Data Mining. A dataset with imbalanced classes is a dataset that the event frequency of certain classes is different from other classes. The imbalanced problem gives a bias in the classification performance. Many researchers have developed algorithms and modified the preprocessing stage to overcome the problem. Hence, this research focused on algorithm comparison between One Class Classification algorithms, which are Elliptic Envelope and Isolation Forest, on solving of the imbalanced data. The results showed that the Elliptic Envelope method performed a better performance compared to the Isolation Forest algorithm by giving 80.28% of recall testing and 80.28% of precision meanwhile Isolation Forest only giving 46.95% of recall testing and 46.95% of precision.

Keywords: *Data mining, Imbalance Datasets, Classification, Elliptic Envelope, Isolation Forest*

Abstrak

Masalah data tidak seimbang merupakan hal yang penting dalam bidang *Data Mining*. *Dataset* dengan kelas yang tak seimbang adalah dataset yang frekuensi kejadian dari kelas tertentu sangat jauh berbeda dengan kelas yang lain. Masalah ketidakseimbangan ini akan memberi bias terhadap performa pengklasifikasi. Banyak peneliti telah mengkaji baik berupa pengembangan algoritma maupun modifikasi pada tahap preprocessing untuk mengatasi masalah ini. Pada penelitian ini membahas komparasi algoritma *One Class Classification* yaitu *Elliptic Envelope* dan *Isolation Forest* pada data tidak seimbang. Dari penelitian ini, algoritma *Elliptic Envelope* menunjukkan hasil lebih baik dibanding dengan algoritma *Isolation Forest* dengan pengujian *recall* 80.28% dan *precision* 80.28% sedangkan algoritma *Isolation Forest* menunjukkan hasil pengujian *recall* 46.95% dan *precision* 46.95%

Kata kunci: Data Mining, Data Tidak Seimbang, Klasifikasi, *Elliptic Envelope*, *Isolation Forest*

1. PENDAHULUAN

Klasifikasi merupakan bagian penting dari *Data Mining*, klasifikasi akan beroperasi pada data yang diambil dari distribusi yang sama dengan data *training*. Namun, terdapat juga data yang

[*1,2] *Program Studi Ilmu Komputer Universitas Hasanuddin,*
***Email**: supriamir@unhas.ac.id, bagasprasetyo33@gmail.com*

memiliki jumlah kelas yang tidak seimbang antar kelas yang satu dengan yang lainya disebut *imbalanced datasets*. Di dalam *Machine Learning* jika mengunakan pendekatan klasifikasi yang standar, data tidak seimbang menghasilkan *performance* yang kurang bagus. Kinerjanya yang kurang bagus karena klasifikasi mungkin mengabaikan pentingnya kelas minoritas karena perwakilannya dalam dataset tidak cukup kuat.

Penelitian pada data tidak seimbang merupakan salah satu topik yang menantang pada bidang *Machine Learning*. Berbagai solusi yang telah diusulkan para peneliti pada masalah ini seperti pengembangan algoritma maupun modifikasi pada tahap *preprocessing*. Dalam tahap *preprocessing* berfokus pada menyeimbangkan data (*rebalanced dataset*). Banyak cara yang telah ditemukan untuk mengatasi dataset tidak seimbang ini, seperti melakukan *resampling* terhadap data yang ada. *Resampling* adalah teknik mengambil sampel secara berulang dari sampel data asli. Teknik *resampling* terdiri dari *oversampling*, yaitu mengambil sampel berulang kali dari kelas minoritas; dan *undersampling*, yaitu mengambil sampel secara acak dari kelas mayoritas [1].

Masalah klasifikasi satu kelas (*One Class Classification*) menjadi sangat penting dalam bidang *Machine Learning*. Istilah klasifikasi satu kelas telah digunakan oleh *Moya* dan *Hush* pada papernya yang berjudul "*Network constraints and multi-objective optimization for one-class classification*" [4]. Banyak penelitian yang telah dilakukan pada masalah klasifikasi satu kelas dengan aplikasi berbeda, misalnya deteksi outlier/deteksi anomali dan *Novelty detection*.

Dalam deteksi anomali klasifikasi satu kelas bekerja untuk mengidentifikasi item atau peristiwa tidak terduga dalam kumpulan data, yang berbeda dari data normal. Deteksi anomali memiliki dua asumsi dasar yaitu anomali sangat jarang terjadi dan fitur anomali berbeda dari data normal. Hal ini tidak berbeda dengan pada dataset tidak seimbang dimana anomali sama dengan data minoritas pada dataset tidak seimbang.

Noumir dkk yang mengembangkan sebuah algoritma yang dinamakan *One Simple Class Classification* pada masalah klasifikasi [5]. Selanjutnya Burnave dkk menggunakan *One Class SVM* untuk mendeteksi *Malware* [2].

Satu perbedaan mendasar antara klasifikasi satu kelas dan klasifikasi standar adalah bahwa dalam pembelajaran satu kelas, diasumsikan bahwa hanya informasi kelas target yang tersedia. Dengan kata lain, dalam proses pelatihan *classifier*, jumlah kelas data dari kelas target digunakan dan tidak ada informasi tentang rekannya. Batas antara kedua kelas harus diperkirakan dari data dari satu-satunya objek yang tersedia.

Menerapkan klasifikasi satu kelas untuk data yang tidak seimbang adalah arah penelitian yang jarang dilakukan, meskipun beberapa pekerjaan telah dilakukan. 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. Pada penelitian ini dilakukan perbandingan dua algoritma klasifikasi satu kelas yaitu *Elliptic Envelope* dan *Isolation Forest* untuk masalah klasifikasi pada data tidak seimbang.

2. STUDI LITERATUR

2.1. Data Tidak Seimbang

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. 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 kelas yang tidak seimbang. Istilah ini berlaku ketika kelas dari dataset tersebut bersifat kategorik diskrit. Dataset dengan kelas yang

tidak 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 [3].

2.2 *Confusion Matrix*

Di dalam *Machine Learning*, mengukur kinerja atau performa dari suatu model merupakan 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* (TP), jika kelas positif tersebut diprediksi negatif (salah) disebut *false negative* (FN). Kelas negatif yang berhasil diprediksi negatif (benar) disebut *true negative* (TN), dan kelas negatif yang diprediksi positif disebut *false positive* (FP). Jumlah dari kasus-kasus tersebut direpresentasikan dalam suatu tabel 2.1 kontingensi yang disebut *confusion matrix.*

Tabel 2.1 *Confusion Matrix*

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

Akurasi *(Accuracy)* 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:

Presisi *(Precision)* 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:

$$Precision = \frac{TP}{(TP + FP)} \quad 1.1$$

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:

$$Recall = \frac{TP}{(TP + FN)} \quad 2.2$$

F1-Score adalah harmonic mean antara precision dan recall:

$$F1 = 2 * \frac{Precision * Recall}{(Precision + Recall)} \quad 2.3$$

2.3 Algoritma Klasifikasi Satu Kelas (*One Class Classification*)

2.3.1 *Elliptic Envelope*

Elliptic Envelope adalah fungsi yang mencoba mencari tahu parameter kunci dari distribusi umum data dengan mengasumsikan bahwa seluruh data adalah ekspresi dari distribusi normal/*Gaussian multivarian* yang mendasarinya. Metode ini mencoba menemukan batas *elips* sebagian besar data, data yang berada di luar *elips* akan diklasifikasi sebagai anomali. *Elliptic Envelope* menggunakan *FAST-minimum covariance determinant* untuk memperkirakan ukuran dan bentuk elips [6].

FAST-minimum covariance determinat memilih subsampel data yang tidak tumpang tindih dan menghitung *mean* (μ), dan kovarians matriks di setiap dimensi fitur dalam subsample. Jarak *Mahalobis* (d_{MH}) mengurutkan setiap subsample dan data dari yang terkecil ke yang terbesar. Jarak *Mahalobis* didefinisikan sebagai

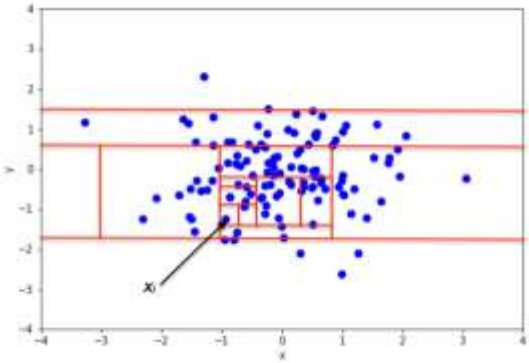
$$d_{MH} = \sqrt{(x-\mu)^T C^{-1}(x-\mu)} \quad 2.4$$

Jarak *Euclidean* akan dikurang jika matriks kovarian adalah matriks identitas. Dan jarak *Euclidean* akan dinormalkan jika matriks kovarian diagonal. Secara sederhana, jarak *Mahalobis* mengukur berapa banyak sigma titik data dari distribusi mean.

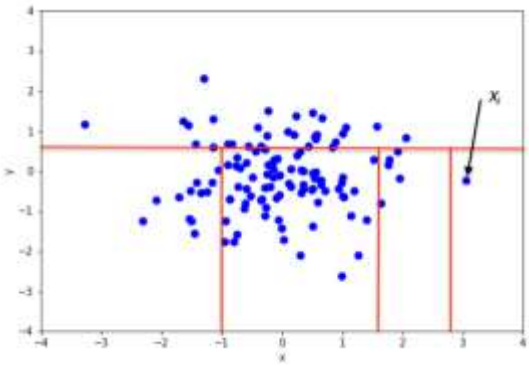
Metode *FAST-minimum covariance determinate* melanjutkan dengan memilih subsampel dari sampel asli, dengan nilai d_{MH} yang kecil. *Mean*,kovarian, dan nilai d_{MH} dari subsampel dihitung kembali. Prosedur ini diulang sampai determinasi matriks kovarian menyatu. Matriks kovarians dengan determinasi terkecil dari semua subsampel membentuk sebuah *elips* yang mencakup sebagian kecil data asli. Data dalam elips diberi label sebagai "*inlier*"dan data diluar elips diberi label "*outlier*" atau anomali yang kemudian dapat dihilangkan.

2.3.2 *Isolation Forest*

Algoritma *Isolation Forest* pada awalnya diusulkan oleh Liu dkk [7]. Algoritma ini berbeda dari kebanyakan algoritma klasifikasi satu kelas lainnya, yang secara eksplisit mengenali anomali/*outliers* yang dalam penelitian ini disebut kelas minoritas daripada membuat profil data normal yang dalam penelitian ini disebut kelas mayoritas. Secara teoritis, data normal terjadi lebih sering daripada *outliers*. Selain itu pengamatan untuk data normal dan anomali berbeda satu sama lain dalam hal nilai. Dalam banyak kasus anomali terletak lebih jauh dari titik data biasa dalam satu ruang fitur (gambar 2.2). Berdasarkan fakta ini, titik data anomali memerlukan lebih sedikit partisi untuk diidentifikasi daripada data normal (gambar 2.1).



Gambar 2.1. Contoh mengisolasi titk data normal



Gambar 2.2. Contoh mengisolasi titik data anomali

Berdasarkan itu data anomali dalam dataset ada kecenderungan lebih mudah untuk dipisahkan dari sisa sampel, dibandingkan dengan data normal. Dalam rangka untuk mengisolasi titik data, algoritma *Isolation Forest* secara rekursif menghasilkan partisi pada sampel dengan memilih secara acak atribut dan kemudian secara acak memilih nilai terpisah untuk atribut, antara nilai minimum dan maksimum yang diizinkan untuk atribut tersebut.

Isolation Forest menggunakan skor anomali untuk membuat keputusan. Skor anomali s dari *instance* x dapat didefinisikan sebagai:

$$s(x,n) = 2^{\frac{E(h(x))}{c(n)}} \quad 2.5$$

Dimana $h(x)$ adalah panjang jalur titik x, $E(h(x))$ adalah rata-rata $h(x)$ dari kumpulan pohon isolasi. $c(n)$ adalah rata-rata panjang jalur pencarian yang tidak berhasil di pohon pencarian biner. n adalah jumlah eksternal *node*.

3. METODE PENELITIAN

3.1. Deskripsi Data

Dataset yang digunakan adalah *Credit Fraud* Dataset yang diperoleh dari website resmi Kagle (https://www.kaggle.com/mlgulb/creditcardfraud) 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 hanya memiliki 2 kelas yaitu transaksi normal (mayoritas) dan transaksi fraud (minoritas).

3.2. Tahapan Penelitian

Penelitian ini melewati beberapa tahap, yaitu :

1. Tahap eksplorasi dan *preprocessing* data, peneliti mencoba menguraikan karakteristik-karakteristik setiap dataset sebagai informasi untuk mengambil keputusan pada tahap *preprocessing*. Peneliti akan mengidentifikasi masalah-masalah yang terdapat pada dataset tersebut kemudian mengambil pendekatan untuk menyelesaikan masalah yang terkait. *Preprocessing* yang akan dilakukan pada data antara lain:
    a) *Cleaning*: pada tahap ini dataset dibersihkan dari *missing value*, menghilangkan observasi yang tidak diinginkan.
    b) Normalisasi: Normalisasi yang akan digunakan yaitu *standarscaler* yang terdapat pada library sklearn yang berfungsi mengubah data sehingga distribusi data memiliki rata-rata 0 dan standar deviasi 1.
    c) Membagi data: Data akan dibagi menjadi 2 yaitu data *training* dan data *testing*.
2. Tahap model *tuning* dan *fitting*, peneliti mencari parameter-parameter terbaik untuk model yang digunakan berdasarkan hasil eksplorasi data dan *trial* dan *error* untuk mendapatkan hasil terbaik. *Tuning* juga dilakukan terhadap beberapa teknik deteksi *outlier* yang membutuhkan parameter. Kemudian model akan memberi hasil prediksi yang akan dianalisis pada tahap selanjutnya. Algoritma klasifikasi satu kelas yang digunakan pada penelitian ini adalah *Elliptic Envelope* dan *Isolation Forest*.
3. Tahap analisis hasil, peneliti akan merangkum hasil yang diperoleh dari algoritma-algoritma yang digunakan ke dalam bentuk tabel dan diagram, kemudian menyimpulkan hasilnya sebagai output dari penelitian ini.

4. HASIL PEMBAHASAN

4.1. *Preprocessing*

Pada tahap *preprocessing* melakukan proses normalisasi pada dataset menggunakan *standarscaler* yang akan mengubah data sehingga distribusi data memiliki rata-rata 0 dan standar deviasi 1. Proses ini bertujuan untuk memperkecil skala dari dataset tanpa mempengaruhi distribusinya agar memudahkan model belajar lebih cepat dan akurat. Pada tabel 4.1 nilai *mean* 2 dari 30 atribut data sebelum proses normalisasi dan setelah normalisasi.

Tabel 4.1. Mean Data Atribut

Atribut	Mean sebelum normalisasi	Mean setelah normalisasi
Amount	88.291	2.91
Time	94838.20	-3.06

Setelah itu dilakukan pemberian label 0 untuk transaksi normal dan label 1 untuk transaksi *fraud*. Kemudian dataset dibagi menjadi 2 bagian yaitu data *training* dan data *testing* dengan rasio 70:30. Data *training* berisi 209315 data transaksi normal dan tidak ada satupun transaksi *fraud* dalam data *training*. Data *testing* berisi 75492 data yang merupakan campuran transaksi normal dan transaksi *fraud*.

4.2. *Tuning* dan *Fitting* pada *Elliptic Envelope*

Proses dalam menentukan parameter pada metode disebut proses *tuning*. Parameter untuk algoritma *Elliptic Envelope* adalah *support fraction* dan untuk nilai awal *support fraction* yang digunakan pada penelitian ini adalah 0.994.

Model akan dibuat berdasarkan data *training* dan hasil model tersebut akan melalui proses data *testing* atau disebut *fitting*. Kemudian akan dilakukan perulangan pada proses *tuning* dan *fitting* hingga parameter terbaik didapatkan berdasarkan nilai skor F1. Pada tabel 4.2 adalah hasil pencarian parameter terbaik untuk metode *Elliptic Envelope*.

Tabel 4.2. *Support Fraction* berdasarkan skor F1

Support Fraction	F1
0.95	80.28%
0.96	80.28%
0.97	80.28%
0.98	80.28%
0.99	80.08%

Diperoleh nilai 0.95 sampai 0.98 memiliki hasil skor F1 yang sama. Jadi nilai *support fraction* yang akan digunakan pada algoritma *Elliptic Envelope* adalah 0.95.

4.3. *Tuning* dan *Fitting* pada *Isolation Forest*

Parameter untuk algoritma *Isolation Forest* adalah *num estimator* dan untuk nilai awal *num estimator* yang digunakan pada penelitian ini adalah 1050. Setelah itu model akan dibuat berdasarkan data *training* dan hasil model tersebut akan di *test* pada data *testing* atau disebut *fitting*. Pada tabel 4.3 adalah hasil pencarian parameter terbaik untuk algoritma *Isolation Forest*.

Tabel 4.3. *Num estimator* berdasarkan skor F1

Num estimator	F1
900	46.34%
950	46.54%
1000	46.95%
1050	46.95%
1100	46.95%

Didapatkan nilai 1000 sampai 1100 memiliki hasil skor F1 terbaik. Sehingga nilai *Num estimator* yang akan digunakan pada algoritma *Isolation Forest* adalah 1000.

4.4. Hasil perbandingan

Penelitian ini membandingkan dua algoritma yaitu *Elliptic Envelope* dan *Isolation Forest* pada masalah klasifikasi pada data tidak seimbang. Penelitian ini membandingkan *recall*, *precision* dan *F1* dari hasil klasifikasi algoritma tersebut. Nilai yang akan di perbandingkan merupakan parameter terbaik yang didapat berdasarkan skor F1 masing-masing metode seperti yang ditunjukkan pada tabel 4.4

Tabel 4.4. Hasil Perbandingan *Recall, Precision,* dan skor F1

Metode	*Recall*	*Precision*	F1
Elliptic Envelope	80.28%	80.28%	80.28%
Isolation Forest	46.95%	46.95%	46.95%

5. KESIMPULAN

Dari hasil percobaan terbukti bahwa algoritma klasifikasi satu kelas yang lebih baik kinerjanya pada klasifikasi dataset tidak seimbang adalah *Elliptic Envelope* dibandingkan *Isolation Forest*. Algoritma *Elliptic Envelope* menunjukkan hasil pengujian *recall* 80.28% dan *precision* 80.28% sedangkan *Isolation Forest* menunjukkan hasil pengujian *recall* 46.95% dan *precision* 46.95%

DAFTAR PUSTAKA

[1] Burnaev, E., Erofeev, P. and Papanov, A., (2015). Influence of resampling on accuracy of imbalanced classification. *Eighth International Conference on Machine Vision (ICMV 2015)*, 9875(December), p.987521.

[2] Burnaev, E. and Smolyakov, D., (2016). One-Class SVM with Privileged Information and Its Application to Malware Detection. *IEEE International Conference on Data Mining Workshops*, ICDMW, 0, pp.273–280.

[3] Kang, S. and Ramamohanarao, K., (2014). A robust classifier for imbalanced datasets. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8443 LNAI(PART 1), pp.212–223.

[4] Moya, M.M. and Hush, D.R., (1996). Network constraints and multi-objective optimization for one-class classification. *Neural Networks*, 9(3), pp.463–474.

[5] Noumir, Z., Honeine, P. and Richard, C., (2012). On simple one-class classification methods. *IEEE International Symposium on Information Theory - Proceedings, (October)*, pp.2022–2026.

[6] Rousseeuw, P. and Driessen, K., (1999). A Fast Algorithm for the Minimum Covariance. *Technometrics*, 41(3), pp.212–223.

[7] Tony Liu, F., Ming Ting, K. and Zhou, Z.-H., (2008). Isolation Forest ICDM08. *Icdm*.