PENINGKATAN PERFORMA ALGORITMAK-NEAREST NEIGBORD DALAM KELASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN METODE SPIDER-2

Resianta Perangin-angin[1], Sanco Simanullang[2] , Darwis Robinson Manalu[3]

[1]Komputerisasi Akuntansi, Universitas Methodist Indonesia
[2]Pendidikan Teknologi Informasi, Universitas Methodist Indonesia
[3]Sistem Informasi, Universitas Methodist Indonesia

[1]resianta88@gmail.com, [2]sancosimanullang@gmail.com,[3]manaludawis@gmail.com

ABSTRACT

Class imbalance has become an ongoing problem in the field of Machine Learning and Classification. The group of data classes that are less known as the minority group, the other data class group is called the majority group (majority). In essence real data, data that is mined directly from the database is unbalanced. This condition makes it difficult for the classification method to perform generalization functions in the machine learning process. Almost all classification algorithms such as Naive Bayes, Decision Tree, K-Nearest Neighbor and others show very poor performance when working on data with highly unbalanced classes. The classification methods mentioned above are not equipped with the ability to deal with class imbalance problems. Many data processing methods are often used in cases of data imbalance, in this case research will be carried out using the Spider2 method. In this study, the Ecoli dataset was used, while for this study, 5 (five) different Ecoli datasets were used for each dataset for the level of data imbalance. After testing datasets with different levels of Inbalancing Ratio (IR), starting from the smallest 1.86 to 15.80, the results that explain that the KNN algorithm can improve its performance even better in terms of unbalanced data classification by adding the SPIDER- method 2 as a tool in dataset processing. In the 5 trials, the performance of the KNN algorithm can increase GM by 5.81% and FM 14.47% by adding the SPIDER-2 method to KNN.

Kata Kunci: Data Mining, SPIDER-2, KNN, Unbalancing Data

I. PENDAHULUAN

Ketidakseimbangan kelas telah menjadi masalah yang sedang berlangsung bidang *Machine Learning* dan Klasifikasi. Ketika ketersediaan data awal tumbuh pada tingkat eksponensial, ada banyak cara di bidang teknik dan sains untuk membangun sistem pembelajaran dan kecerdasan buatan untuk memilah-milah data dalam jumlah besar[8]. Kelompok kelas data yang lebih sedikit dikenal dengan kelompok minoritas (*minority*), kelompok kelas data yang lainnya disebut dengan kelompok mayoritas (*mayority*)[2][5][8]. Pada hakekatnya data yang ditambang langsung dari database adalah tidak seimbang. Kondisi tersebut menyulitkan metode klasifikasi dalam melakukan fungsi generalisasi pada proses machine learning. 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. Metode-metode klasifikasi yang telah disebutkan di atas tidak dilengkapi dengan kemampuan untuk menangani masalah ketidak seimbangan kelas [13]. Dalam masalah klasifikasi dengan data kelas tidak seimbang hal ini menjadi masalah yang *urgent* dalam bidang ilmu *machine learning* dan data mining, sebagai contoh dalam permasalahan dunia medis.(Kothandan, 2015), dalam permasalah klasifikasi text. (Wu et al., 2014), dan juga untuk permasalah di dunia sosial media.[9]. Hampir semua algoritma klasifikasi yang populer digunakan, akan tidak maksimal dalam mengklasifikan data yang memiliki tingkat ketidakseimbangan data yang besar.[1][6][10]. Perbedaan ini merupakan suatu indikator performa klasifikasi yang buruk. Dalam beberapa kasus, dimana ternyata kelas minoritas justru lebih penting untuk diidentifikasi daripada kelas mayoritas[12]. Sebagai contoh untuk kasus transaksi menggunakan kartu kredit, dimana kebanyakan status transaksi adalah transaksi yang normal, hanya sedikit kasus yang dapat ditemukan dimana terjadi transaksi yang fraud. Namun Demikian, keberadaan transaksi yang fraud jauh lebih penting untuk di identifikasi daripada transaksi yang normal meskipun jumlah kasusnya jauh lebih sedikit[3][15]. Banyak methode prosesing data yang sering digunakan dalam kasus ketidakseimbangan data, dalam kasus ini akan dilakukan penelitian dengan menggunakan metode Spider2. Dalam penelitian[11]. Digunakan dataset Ecoli, sedangkan untuk penelitian ini akan dilakukan dengan menggunakan 5 (lima) dataset Ecoli yang berbeda setiap dataset untuk tingkat ketidakseimbangan datanya.

II. METODE SPIDER-2

Dimana metode ini Metode ini terdiri dari dua fase yang sesuai dengan preprocessing kelas mayoritas dan minoritas. Pada fase pertama, mengidentifikasi karakteristik contoh darikelas mayoritas, dan itu menghilangkan atau memberi label ulang contoh berisik

dari yang satu ini. Pada fase kedua, mengidentifikasi karakteristik contoh dari kelas minoritas dengan mempertimbangkan perubahan diperkenalkan pada fase pertama. Kemudian, contoh bising dari kelas minoritas diperkuat dengan mereplikasinya[11].

Konsep dalam bentuk narasi dilengkapi dengan formula meotde Spider-2

Metode K-Nearest Neigbord

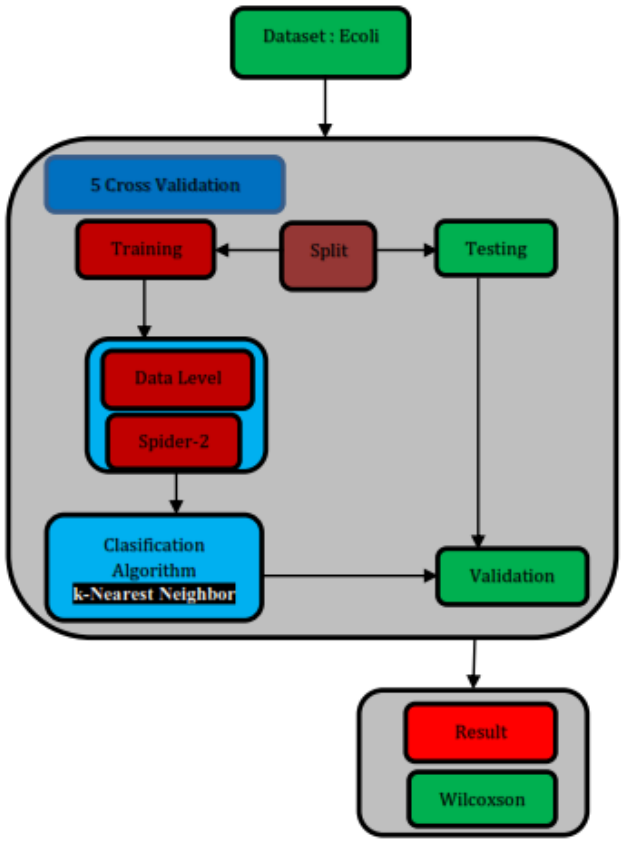
Dalam Metode k-Nearest Neighbor (k-NN) dimana metode klasifikasi yang klasik yang paling sederhana. Metode k-NN sering juga disebut dengan Instance Based Learning, k-NN melakukan klasifikasi terhadap objek berdasarkan jarak antara objek tersebut dengan objek lain. Metode k-NN menggunakan prinsip ketetanggaan (neighbor) untuk memprediksi kelas yang baru. Jumlah tetangga yang dipakai adalah sebanyak k tetangga.[13].

III. METODOLOGI

Dalam penelitian ini akan dilakukan pengukuran performa dari metode yang digunakan yakni SPIDER-2 + KNN dan KNN saja. Untuk dataset yang digunakan menggunakan dataset 5 jenis dengan tingkat ketidakseimbangan yang berbeda-beda, dan nilai K=1 untuk semua ujicoba dataset.

Kerangka Penelitian

Penelitian ini dilakukan dengan mengusulkan sebuah model, dengan mengimplementasikan sebuah metode pemrosesan data dengan menggunakan software keel dan mensintesis data latih. Algoritma resampling yang digunakan adalah SPIDER-2 sedangkan Algoritma pengklasifikasi yang digunakan adalah K-NN. Validasi pada pengukuran kinerja digunakan 5- fold cross validation. Hasil pengukuran dianalisa menggunakan wilcoxon. Kerangka kerja model yang diusulkan ditunjukkan pada Gambar 1.



Gambar 1. Kerangka Penelitian

IV. HASIL DAN PEMBAHASAN

Dataset

Penelitian ini menggunakan 5 Jenis dataset yang bersumber dari UCI repository. Dataset ini memiliki tingkat ketidakseimbangan data atau Imbalance Ratio (IR) yang berbeda-beda. Yakni dataset Ecoli dengan tingkat IR masing-masing : 1.86, 3.36, 5.46, 8.60 dan 15.80. dan attribut : Mcg, Gvh, Lip, Chg, Aac, Alm1, Alm2.

Prosedur SPIDER-2+KNN

Untuk prosedur dari algoritma KNN + Metode Spider-2, ada beberapa tahapan yang akan dilakukan, untuk melihat performa dari algoritma KNN + SPIDER-2.

1. Partisi dataset secara acak menjadi 5 bagian dengan skema 5-fold cross validation
2. Menerapkan penanganan kelas data tidak seimbang pada ROS dan RUS sebanyak 2 (dua) kali pada data latih :
   a. Menentukan nilai tetangga dengan k = 5
   b. Menghitung jarak antar data kelas minoritas dengan metode eucledian
   c. Melakukan perhitungan untuk membangkitkan data buatan (*syntetic*)
3. Menerapkan k-nearest neighbor untuk mengklasifikasi data uji :
   a. Menentukan nilai tetangga dengan k = 1

b. Menghitung jarak antar data kelas minoritas dengan metode eucledian

4. Melihat performa dari algoritma KNN dalam klasifikasi data tidak seimbang dengan menerapkan metode SPIDER-2 sebagai metode prosesing datanya. Dimana performa klasifikasi yang diterapkan adalah G-Mean dan F-Mean.

Hasil Pengujian

Tabel 1: Hasil Pengujian Dataset Ecoli IR=1.86

Performa	**KNN**	**Spider+KNN**
TP	139	139
TN	74	74
FP	4	4
FN	3	3
Acc	0.97	0.97
TP	139	139
Recal	0.98	0.98
Speci	0.95	0.95
Prec	0.97	0.97
GM	0.96	0.96

Dari tabel pengujian pertama dengan tingkat IR sebesar 1.86 dimana dataset ini berarti hanya memiliki tingkat ketidakseimbangan data sangat kecil, terlihat KNN sangat baik dalam hal performa hampir tidak ada perubahan yang signifikan walau ditambah dengan metode SPIDER-2.

Tabel 2: Hasil Pengujian Dataset Ecoli IR=3.36

Performa	**KNN**	**Spider+KNN**
TP	56	62
TN	239	237
FP	21	15
FN	20	22
Acc	0.88	0.89
Recal	0.74	0.74
Speci	0.92	0.94
Prec	0.73	0.81
GM	0.82	0.83
F-M	0.73	0.77

Dari hasil pengujian yang kedua dengan dataset yang memiliki tingkat IR sebesar 3.36 disini terlihat dengan menambahkan metode SPIDER-2 maka performa dari KNN sudah lebih baik.

Tabel 3: Hasil Pengujian Dataset Ecoli IR=5.46

Performa	**KNN**	**Spider+KNN**
TP	43	46
TN	272	271
FP	9	6
FN	12	13
Acc	0.94	0.95
Recal	0.79	0.79
Speci	0.97	0.98
Prec	0.82	0.89
GM	0.87	0.88
F-M	0.80	0.84

Dari hasil pengujian yang ketiga dengan dataset yang memiliki tingkat IR sebesar 5.46 disini terlihat dengan menambahkan metode SPIDER-2 maka performa dari KNN sudah lebih baik.

Tabel 4 : Hasil Pengujian Dataset Ecoli IR=8.60

Performa	**KNN**	**Spider+KNN**
TP	18	16
TN	284	314
FP	17	4
FN	17	2
Acc	0.90	0.98
Recal	0.51	0.88
Speci	0.94	0.99
Prec	0.51	0.79
GM	0.69	0.93
F-M	0.51	0.83

Tabel 5: Hasil Pengujian Dataset Ecoli IR=15.80

Performa	**KNN**	**Spider+KNN**
TP	15	16
TN	314	314
FP	5	4
FN	2	2
Acc	0.98	0.98
Recal	0.87	0.88
Speci	0.98	0.99
Prec	0.75	0.79
GM	0.93	0.93
F-M	0.81	0.83

Tabel 6: Hasil Pengujian Dataset Ecoli IR=15.80

Performa	KNN	Spider+KNN
TP	54	56
TN	237	242
FP	11	7
FN	11	9
Acc	0.93	0.95
Recal	0.78	0.85
Speci	0.95	0.97
Prec	0.76	0.85
GM	0.86	0.91
F-M	0.76	0.85

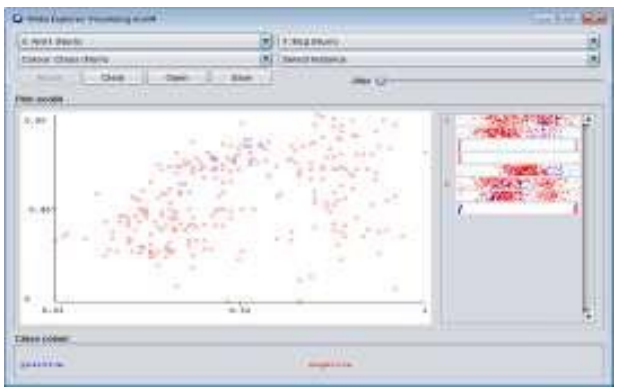
Tabel 7: Hasil Performa dari Algoritma KNN

Performa	KNN	Spider+KNN
TP	54	56
TN	237	242
FP	11	7
FN	11	9
Acc	0.93	0.95
Recal	0.78	0.85
Speci	0.95	0.97
Prec	0.76	0.85
GM	**0.86**	**0.91**
F-M	**0.76**	**0.85**

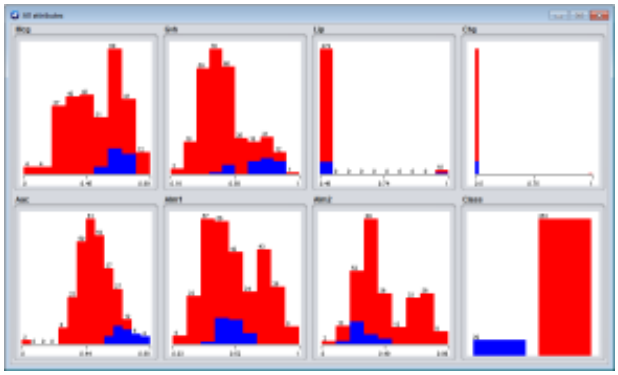
Dari hasil semua pengujian didapat hasil Gm dan Fm dari algoritma KNN + SPIDER-2 akan lebih signifikan pengaruhnya terhadap data dengan tingkat IR yang tinggi dimana semakin tinggi IR akan semakin baik pengaruhnya. Dari hasil ujicoba yang dilakukan maka dapat dilihat kemampuan metode SPIDER-2 dalam meningkatkan performa algoritma KNN dalam klasifikasi data tidak seimbang.

Visualisasi Attribut Dataset

Dalam visualisasi terlihat penyebaran dari atribut dataset Ecoli yang memilih 2 kelas yakni Positive dan Negative.



Gambar 2. Penyebaran Attribut Dataset



Gambar 3. Hasil Visualisasi Seluruh Attribut

V. KESIMPULAN

Setelah melakukan uji coba dataset dengan tingat Inbalancing Ratio (IR) yang berbeda-beda, mulai dari yang terkecil 1.86 sampai dengan 15.80, maka di dapat kesimpulan yang menerangkan bahwa algoritma KNN bisa bertambah performanya lebih baik lagi dalam hal klasifikasi data tidak seimbang dengan menambahkan metode SPIDER-2 sebagai alat bantu dalam pemrosesan dataset. Dalam ujicoba yang dilakukan sebanyak 5 kali percobaan performa algoritma KNN dapat meningkatkan GM sebesar 5.81% dan FM 14.47% dengan menambahkan metode SPIDER-2 kedalam KNN.

DAFTAR PUSTAKA

[1]. Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (n.d.). *Classification with class imbalance problem: A review*. 31.

[2]. Bria, A., Karssemeijer, N., & Tortorella, F. (2014). Learning from unbalanced data: A cascade-based approach for detecting clustered microcalcifications. *Medical Image Analysis*, *18*(2), 241–252. https://doi.org/10.1016/j.media.2013.10.014

[3]. Cordón, I., García, S., Fernández, A., & Herrera, F. (2018). Imbalance: Oversampling algorithms for imbalanced classification in R. *Knowledge-Based Systems*, *161*, 329–341. https://doi.org/10.1016/j.knosys.2018.07. 035

[4]. Department of Biological Sciences, BITS PILANI K K Birla Goa Campus, Zuarinagar, Vasco Da Gama, India, & Kothandan, R. (2015). Handling class imbalance problem in miRNA dataset associated with cancer. *Bioinformation*, *11*(1), 6–10. https://doi.org/10.6026/97320630011006

[5]. Farquad, M. A. H., & Bose, I. (2012). Preprocessing unbalanced data using support vector machine. *Decision Support Systems*, *53*(1), 226–233. https://doi.org/10.1016/j.dss.2012.01.016

[6]. Juan Carbajal-Hernández, J., Sánchez-Fernández, L. P., Hernández-Bautista, I., Medel-Juárez, J. de J., & Sánchez-Pérez, L. A. (2016). Classification of unbalance and misalignment in induction motors using orbital analysis and associative memories. *Neurocomputing*, *175*, 838–850. https://doi.org/10.1016/j.neucom.2015.06.094

[7]. Kothandan, R. (2015). Handling class imbalance problem in miRNA dataset associated with cancer. *Bioinformation*, *11*(1), 6–10. https://doi.org/10.6026/97320630011006

[8]. Lee, J., Wu, Y., & Kim, H. (2015). Unbalanced data classification using support vector machines with active learning on scleroderma lung disease patterns. *Journal of Applied Statistics*, *42*(3), 676–689. https://doi.org/10.1080/02664763.2014.978270

[9]. Li, C., & Liu, S. (2018). A comparative study of the class imbalance problem in Twitter spam detection. *Concurrency and*

Computation: Practice and Experience, *30*(5), e4281. https://doi.org/10.1002/cpe.4281

[10]. Maldonado, S., López, J., & Vairetti, C. (2019). An alternative SMOTE oversampling strategy for high-dimensional datasets. *Applied Soft Computing*, *76*, 380–389. https://doi.org/10.1016/j.asoc.2018.12.024

[11]. Napierała, K., Stefanowski, J., & Wilk, S. (2010). Learning from Imbalanced Data in Presence of Noisy and Borderline Examples. In M. Szczuka, M. Kryszkiewicz, S. Ramanna, R. Jensen, & Q. Hu (Eds.), *Rough Sets and Current Trends in Computing* (Vol. 6086, pp. 158–167). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-13529-3_18

[12]. Qiong, G. (2016). *An Improved SMOTE Algorithm Based on Genetic Algorithm for Imbalanced*. *14*(2), 12.

[13]. Siringoringo, R. (2018). *Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Smote Dan K-Nearest Neighbor*. 6.

[14]. Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). CLASSIFICATION OF IMBALANCED DATA: A REVIEW. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(04), 687–719. https://doi.org/10.1142/S0218001409007326

[15]. D. R. Manalu, M. Zarlis, H. Mawengkang, and O. S. Sitompul, "Forest Fire Prediction in Northern Sumatera using Support Vector Machine Based on the Fire Weather Index," AIRCC Publ. Corp., vol. 10, no. 19, pp. 187– 196, 2020, doi: 10.5121/csit.2020.101915.

[16]. Wu, Q., Ye, Y., Zhang, H., Ng, M. K., & Ho, S.- S. (2014). ForesTexter: An efficient random forest algorithm for imbalanced text categorization. *Knowledge-Based Systems*, *67*, 105–116. https://doi.org/10.1016/j.knosys.2014.06.004