PENGGUNAAN K-MEANS CLUSTERING UNTUK MENGATASI IMBALANCE DATA DENGAN ELM (EXTREME LEARNING MACHINE) SEBAGAI CLASSIFIER

by Umi Mahdiyah

Submission date: 09-Apr-2021 02:30AM (UTC+0700)
Submission ID: 1553917508
File name: ATA_DENGAN_ELM_EXTREME_LEARNING_MACHINE_SEBAGAI_CLASSIFIER.docx (95.53K)
Word count: 1260
Character count: 7490

PENGGUNAAN K-MEANS CLUSTERING UNTUK MENGATASI *IMBALANCE DATA* DENGAN ELM (*EXTREME LEARNING MACHINE*) SEBAGAI *CLASSIFIER*

Umi Mahdiyah[1,a)], Nalsa Cintya Resti[2)], Patmi Kasih[1)]
[1)]*Universitas Nusantara PGRI Kediri*
[2)]*Institut Agama Islam Negeri Kediri*

[a)]*umimahdiyah@gmail.com*

Abstrak

Masalah *imbalanced data* (data yang tidak seimbang) selama ini masih menjadi masalah yang cukup penting pada masalah klasifikasi. Data real pada kehidupan sehari-hari umumnya bersifat imbalance, sehingga untuk melakukan klasifikasi butuh proses yang optimal untuk mendapatkan hasil yang optimal pula. Karena data yang tidak seimbang menyebabkan proses klasifikasi menjadi tidak optimal. Pada artikel ini dilakukan proses *undersampling* dengan memanfaatkan algoritma clustering yaitu K-Means. Selanjutnya, data yang sudah dilakukan proses *undersampling* dimasukkan pada proses klasifikasi menggunakan *Extreme Learning Machine*. Data yang digunakan merupakan *benchmark data set* yang bersifat *imbalance*. Hasil yang didapatakan dalam penelitian ini cukup baik, dilihat dari nilai rata-rata *precision, recall*, dan akurasinya.

Kata kunci: *Imbalanced Data*, Undersampling, K-Means, ELM

Pendahuluan

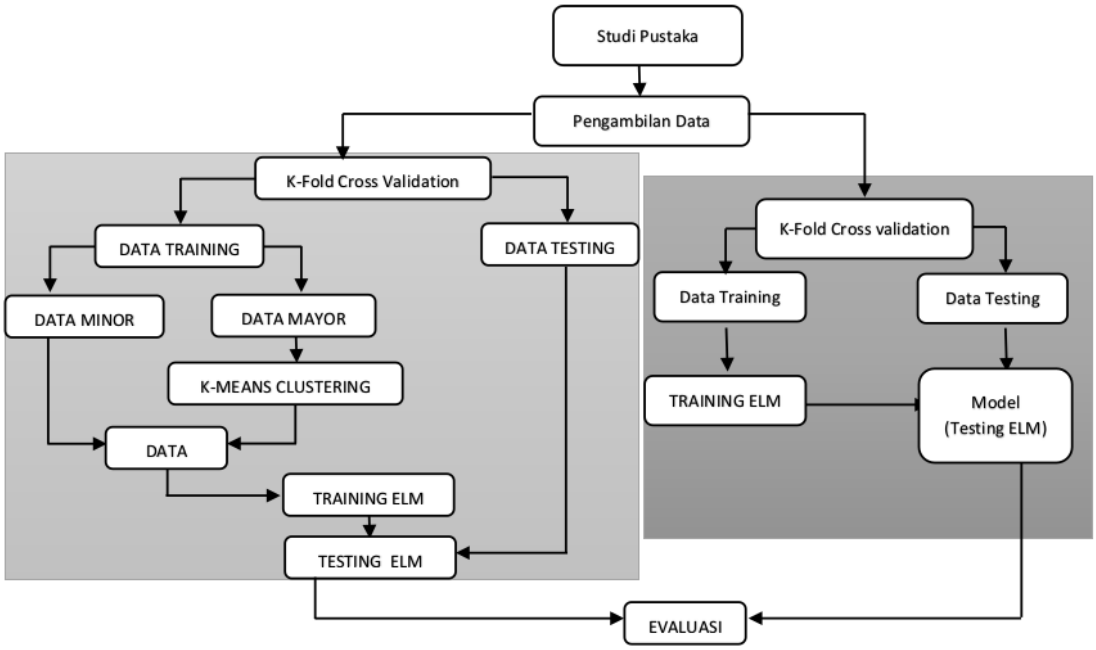
Imbalanced data merupakan salah satu permasalahan yang cukup krusial dalam proses klasifikasi. Karena dengan data yang berkarakter *imbalanced* atau tidak seimbang maka hal tersebut dapat mempengaruhi performansi dari suatu algoritma klasifikasi. Akurasi akan sangat baik untuk kelas mayoritas, akan tetapi akan buruk untuk data minoritas. Permasalahan ketidakseimbangan kelas dapat ditangani dengan 2 pendekatan, yaitu pendekatan level data dan pendekatan level algoritma [1].

Saat ini penanganan masalah imbalanced data sudah cukup banyak dilakukan, diantaranya menggunakan *undersampling* maupun *oversampling* serta pengembangannya [2]. Dalam penelitian ini undersampling dipilih untuk menyelesaikan masalah *imbalanced data* dangan tujuan supaya lebih ringan proses komputasinya, selain itu supaya waktu yang dibutuhkan bisa lebih cepat. K-Means merupakan salah satu algoritma clustering yang sering digunakan karena performansinya yang bagus. Dalam penelitian ini *K-means clustering* digunakan dalam proses *sampling* untuk mengatasi masalah *imbalanced data*. *Extreme Learning Machine* dipilih untuk classifier

Metode

Langkah-langkah dan proses penelitian digambarkan pada Gambar 1. Langkah awal melakukan studi literatur terkait data yang tidak seimbang. Selanjutnya dilakukan pengambilan data pada web http://www.keel.es/. Data yang digunakan dalam penelitian ini dapat dilihat pada Tabel 1, data yang digunakan pada penelitian ini merupakan data biner. Data tersebut kemudian dibagi menjadi 2 bagian yaitu data training dan data testing, dengan mengunakan 5 *fold cross validation*.

Data training adalah data yang digunakan untuk melatih sistem. Dalam penelitian ini data yang digunakan adalah data biner, sehingga data dibagi lagi menjadi data minor dan data mayor. Selanjutnya pada data mayor dilakukan proses sampling dengan menggunakan K-means clustering [3]. Tujuan dari sampling dengan untuk meringkas data berdasarkan kesamaan karakter sehingga data minor dan data mayor tidak memiliki selisih yang sangat banyak.



Gambar 1. Diagram Alir Penelitian

Tabel 1. Data Yang Digunakan Pada Penelitian

No	Nama Data	Banyak Data	
		Data Negatif	Data Positif
1	Pima	150	134
2	Ecoli	129	39
3	Abalone	344	21
4	Glass1	69	38
5	Haberman	113	41
6	Yeast	527	215
7	Spambase	1394	906
8	Cleveland	80	6
9	Dermatology	169	10

Menurut Daniel dan Eko [4], Langkah-langkah algoritma K-Means adalah sebagai berikut:

a. Pilih secara acak k buah data sebagai pusat cluster.

b. Jarak antara data dan pusat cluster dihitung menggunakan Euclidian Distance.

Untuk menghitung jarak semua data ke setiap titik pusat cluster dapat menggunakan teori jarak Euclidean yang dirumuskan sebagai berikut:

$$D_{i,j} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \cdots + (x_{ki} - x_{kj})^2}$$

dimana:

$D(i, j)$ = Jarak data ke i ke pusat cluster j

Xki = Data ke i pada atribut data ke k

Xkj = Titik pusat ke j pada atribut ke k

c. Data ditempatkan dalam cluster yang terdekat, dihitung dari tengah cluster.

d. Pusat *cluster* baru akan ditentukan bila semua data telah ditetapkan dalam cluster

Setelah data sudah di cluster dengan menggunakan *K-means* selanjutnya diambil beberapa bagian data pada setiap cluster disesuaikan dengan perbandingan data tak seimbangnya. Setelah itu data yang sudah diambil dari data mayor digabung lagi dengan data minor, kemudian data gabungan dimasukkan ke dalam algoritma *Extreme Learning Machine*.

Konsep utama dari ELM seperti yang disajikan dalam *paper* Huang adalah sebagai berikut:

Diberikan *training* set $\aleph = \{(x_j, t_j) | x_j \in \boldsymbol{R}^{n \times m}, t_j \in \boldsymbol{R}^n,\ j \in [1, N]\}$, fungsi aktivasi $g(x)$, dan bilangan *hidden node* $\tilde{N}$

Step 1: masukkan secara random bobot w_i dan bias $b_i, i \in [1, \tilde{N}]$

Step 2: hitung *output* matriks *hidden layer H*

Step 3: hitung bobot *output* β

$$\beta = H^{\dagger} T$$

dengan $T = [t_1, \ldots \ldots, t_N]^T$,

$H^{\dagger}$ adalah *Generalized Inverse*

$H^{\dagger} = (H^T H)^{-1} H^T$ atau,

$H^{\dagger} = H^T (H H^T)^{-1}$

Setelah dilakukan proses training dan testing, selanjutnya dilakukan evaluasi dengan menggunakan *confusion matriks*. Untuk mengukur performa dari klasifikasi data pengujian *imbalance* dalam hal ini digunakan *confusion matrix*, *precision* dan *recall*, *specificity*, dan *G-mean* [5]. Tabel *confusion matrix* Dapat dilihat pada tabel 2.

Tabel 2. *Confusion Matrix*

		Nilai Sebenarnya	
		True	*False*
Prediksi	*True*	TP *(True Positive)*	FP *(False Positive)*
	False	FN *(False Negative)*	TN *(True Negative)*

Precision adalah presentase dari data yang diprediksi benar oleh *classifier* yang bernilai benar.

$$precision = \frac{TP}{TP + FP}.$$

Recall adalah porsi dari data sampel yang diprediksi benar oleh *classifier*.

$$recall = \frac{TP}{TP + FN}.$$

sensitifity=recall

Geometric mean telah digunakan beberapa peneliti untuk mengevaluasi *classifier* pada dataset yang *imbalanced* . *G-mean* mengindikasikan keseimbangan antara kinerja klasifikasi pada kelas mayoritas dan minoritas. Ukuran *G-mean* diambil berdasarkan s*ensitifity* (akurasi dari data positif) dan s*pecificity* (akurasi data negatif).

$$specificity = 1 - \frac{TP}{FP + TN}$$

$$G - mean = \sqrt{sensitivity \times specificity}$$

Hasil dan Pembahasan

Hasil dari penelitian dapat dilihat pada Tabel 3. Dari tabel tersebut dapat dilihat bahwa setelah dilakukan sampling pada data yang berkarakter *imbalanced* pada saat proses training dapat meningkatkan performansi *Extreme Learning Machine* untuk kasus data yang tak seimbang. Terlihat pada nilai akurasi data ELM lebih tinggi dibandingkan KM-ELM karena pada ELM saat itu akurasi lebih dominan pada data mayoritas, padahal jika dilihat niali precision, recall, specificity, dan g-mean ternyata kurang baik untuk ELM. Setelah dilakukan proses sampling dengan k-means maka dapat dilihat juga pada tabel 3 bahwa nilai setiap performansinya sudah lebih baik. Masing2 rata2 yang didapatkan untuk akurasi, *precision, recall, specificity,* dan *G-mean* adalah 0,78, 0,87, 0,66, 0,91, dan 0,76.

Tabel 3. Tabel Hasil Penelitian

Nama Data	Akurasi		Precision		Recall		Specificity		G-mean	
	ELM	KM-ELM	ELM	KM-ELM	ELM	KM-ELM	ELM	KM-ELM	ELM	KM-ELM
Pima	0,63	0,65	0,48	0,68	0,52	0,63	0,69	0,68	0,60	0,65
Ecoli	0,84	0,92	0,62	0,97	0,82	0,85	0,85	0,98	0,83	0,91
Abalone	0,94	0,86	0,50	1,00	0,10	0,71	0,99	1,00	0,31	0,85
Glass1	0,64	0,54	0,50	0,65	0,21	0,29	0,88	0,82	0,43	0,49
Haberman	0,74	0,67	0,53	0,82	0,24	0,45	0,92	0,89	0,47	0,63
Yeast	0,73	0,71	0,59	0,81	0,20	0,47	0,94	0,91	0,44	0,66
Spambase	0,69	0,73	0,59	0,88	0,66	0,61	0,71	0,89	0,68	0,74
Cleveland	0,93	1,00	NaN	1,00	0	1,00	1,00	1,00	0	1,00
Dermatology	0,96	0,95	0,71	1,00	0,50	0,90	0,99	1,00	0,70	0,95
Rata-rata	**0,79**	**0,78**	**0,56**	**0,87**	**0,36**	**0,66**	**0,89**	**0,91**	**0,50**	**0,76**

*ELM = Extreme Learning Machine

*KM-ELM = sebelum proses ELM data terlebih dahulu dilakukan sampling dengan K-means

Daftar Rujukan

[1] B. Santoso, H. Wijayanto, K. A. Notodiputro dan B. Sartono, "Class imbalanced problems: a review," dalam *Conference Series: Earth and Environmental Science*, -, 2017.

[2] U. Mahdiyah, M. I. Irawan dan E. M. Imah, "Integrating data selection and extreme learning machine for imbalanced data," *Procedia Computer Science* , pp. 221-229, 2015.

[3] P. Arora, D. Deepali dan S. Varshney, "Analysis of K-Means and K-Medoids Algorithm For Big Data," dalam *International Conference on Information Security & Privacy (ICISP2015)*, India, 2016.

[4] D. R. Kaparang dan E. Sediyono, "Penentuan Alih Fungsi Lahan Marginal Menjadi Lahan Pangan Berbasis," *d'Cartesian: Jurnal Matematika dan Aplikasi* , vol. 2, no. 2, pp. 18-25, 2013.

[5] H. He dan E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering* , vol. 21, no. 9, pp. 1263-1284, 2009.

[6] J. Giganti, "Why Teach Problem Solving, Part I: The World Needs Good Problem Solvers!.," *ComMuniCator,* vol. 31, no. 4, pp. 15-16, 2007.

[7] S. Dewiyani, "Improving Students Soft Skills using Thinking Process Profile Based on Personality Types," *International Journal of Evaluation and Research in Education (IJERE),* pp. pp 118-129, 2015.

PENGGUNAAN K-MEANS CLUSTERING UNTUK MENGATASI IMBALANCE DATA DENGAN ELM (EXTREME LEARNING MACHINE) SEBAGAI CLASSIFIER

ORIGINALITY REPORT

18%	13%	7%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Brawijaya Student Paper	2%
2	repository.mercubuana.ac.id Internet Source	2%
3	informa.poltekindonusa.ac.id Internet Source	1%
4	pure-oai.bham.ac.uk Internet Source	1%
5	ejournal.sisfokomtek.org Internet Source	1%
6	Submitted to Liverpool John Moores University Student Paper	1%
7	pt.scribd.com Internet Source	1%
8	Istiqomah Istiqomah, Habudin Habudin. "ANALISIS NILAI-NILAI PENDIDIKAN DALAM	1%

	SENI TARI AHLAN WASAHLAN DAN TARI RAMPAK TERBANG CIOLANG DAERAH BANTEN", Primary : Jurnal Keilmuan dan Kependidikan Dasar, 2019 Publication	
9	eksplora.stikom-bali.ac.id Internet Source	1 %
10	Azmi Musyaffa Fadhilah, M Iwan Wahyuddin, Deny Hidayatullah. "Analisis Faktor yang Mempengaruhi Perokok Beralih ke Produk Alternatif Tembakau (VAPE) menggunakan Metode K-Means Clustering", Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi), 2020 Publication	1 %
11	jurnal.untad.ac.id Internet Source	1 %
12	scitepress.org Internet Source	1 %
13	Archana Pritam Kale, Shefali Sonavane. "PF-FELM: A Robust PCA Feature Selection for Fuzzy Extreme Learning Machine", IEEE Journal of Selected Topics in Signal Processing, 2018 Publication	1 %
14	repository.its.ac.id Internet Source	1 %

15	Agus Nur Khomarudin, Supratman Zakir, Rina Novita, Endrawati, Mohd Zahiri bin Awang Mat, Efmi Maiyana. "K-Mean Clustering Algorithm in Grouping Prospective Scholarship Recipients", Journal of Physics: Conference Series, 2021 Publication	1 %
16	Submitted to Charotar University of Science And Technology Student Paper	1 %
17	journal.ummat.ac.id Internet Source	1 %
18	mass.iain-jember.ac.id Internet Source	1 %
19	pse.litbang.pertanian.go.id Internet Source	1 %

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		