Terbitan online pada laman https://jurnal.pelitabangsa.ac.id/index.php/jpcs

Journal of Practical Computer Science

JPCS

Vol.2 No.2 November 2022 ISSN Media Elektronik: 2809-8137

Diterima 27 Maret 2023 | Direvisi 28 Maret 2023 | Dipublikasikan 30 Maret 2023

Feature Selection Menggunakan Algoritma Meta-Heuristik

Salamet Nur Himawan[1,*], Rendi[2], Nur Budi Nugraha[3]

[1,2,3,*]Jurusan Teknik Informatika, Politeknik Negeri Indramayu

[1]snhimawan@polindra.ac.id, [2]rendi@polindra.ac.id, [3]nurbudinugraha@polindra.ac.id

Abstract

Machine learning requires data to make predictions. Data can have a large number of features. The large number of features can cause machine learning models to overfit, increase model complexity, and high computational costs. Feature selection is one method for optimizing machine learning models. Feature selection reduces the number of features used in the learning process. This research proposes a feature selection method using meta-heuristic algorithms. The machine learning model serves as the objective function for the meta-heuristic algorithm. The objective function is evaluated at each iteration to obtain the most influential features in the model. The machine learning models used are Random Forest, k-Nearest Neighbors, and Support Vector Machine. The meta-heuristic algorithms used are Differential Evolution, Flower Pollination, Grey Wolf, and Whale Optimization. The research shows that using meta-heuristic algorithms can improve the accuracy of machine learning models with fewer features. The Support Vector Machine – Differential Evolution scheme has the highest accuracy and uses the fewest features.

Keyword: *Feature Selection, Machine Learning, Meta-Heuristic.*

Abstrak

Machine learning membutuhkan data dalam melakukan prediksi. Data dapat memiliki jumlah fitur yang sangat banyak. Jumlah fitur yang sangat banyak membuat model *machine learning overfitting*, meningkatkan kompleksitas model dan meningkatkan biaya komputasi. Salah satu metode dalam optimisasi model *machine learning* yaitu *feature selection*. *Feature selection* dapat mengurangi jumlah fitur yang digunakan model dalam proses belajar. Penelitian ini membuat sebuah metode *feature selection* menggunakan algoritma meta-heuristik. Model *machine learning* menjadi fungsi objektif bagi algoritma meta-heuristik. Fungsi objektif dievaluasi pada setiap iterasi, sehingga mendapatkan fitur yang paling berpengaruh pada model. Model *machine learning* yang digunakan adalah Random Forest, k-Nearest Neighbors dan Support Vector Machine. Algoritma meta-heuristik yang digunakan adalah Differential Evolution, Flower Pollination, Grey Wolf, dan Whale Optimization. Penelitian menunjukan bahwa dengan menggunakan algoritma meta-heuristik dapat meningkatkan akurasi dari model *machine learning* dengan jumalah fitur yang lebih sedikit. Skema *Support Vector Machine – Differential Evolution* merupakan skema dengan nilai akurasi tertinggi dan menggunakan fitur paling sedikit.

Kata kunci: Feature Selection, Machine learning, Meta-Heuristik.

PENDAHULUAN

Model *machine learning* telah banyak berkembang dalam memecahkan berbagai masalah. Klasifikasi menjadi salah satu permasalahan yang dapat dipecahkan dengan algoritma *machine learning*. Beberapa algoritma *machine learning* yang banyak digunakan adalah Random Forest (RF), k-Nearest Neighbors (kNN) dan Support Vector Machine (SVM). Diaz-Uriarte menyampaikan bahwa RF dapat menyelesaikan masalah klasifikasi menggunakan micro array data dan menunjukan hasil performa yang baik (Diaz-Uriarte & Andres, 2006). kNN merupakan algoritma yang baik dalam proses klasifikasi, Nababan mengembangkan attribute weighting yang dapat meningkatkan performa dari kNN (Nababan, dkk., 2018). SVM salah satu algoritma yang paling banyak digunakan oleh peneliti, Cervantes dalam surveynya

menunjukan bahwa SVM merupakan algoritma yang sangat popular dalam menyelesaikan permasalahan klasifikasi dan regresi (Cervantes, dkk., 2020).

Optimalisasi algoritma *machine learning* sangat penting, karena dapat meningkatkan performa dari algoritma tersebut. Berikut merupakan beberapa cara dalam optimalisasi algortima *machine learning,* yaitu *feature selection*, *hyperparameter tuning*, *ensemble learning, cross validation* dan *data preprocessing*. Penelitian ini berfokus pada optimalisasi algoritma *machine learning* menggunakan *feature selection*. Algoritma yang digunakan adalah RF, kNN dan SVM.

Feature selection adalah proses identifikasi subset fitur yang paling relevan dari kumpulan fitur yang tersedia untuk meningkatkan kinerja algoritma *machine learning*. Pada penelitian ini feature selection dilakukan menggunakan algoritma meta-heuristik. Algoritma meta-heuristik merupakan algoritma yang terinspirasi oleh fenomena alamiah seperti evolusi (*evolutionary based*), perilaku kawanan (*swarm based*), dan seleksi alam. Algoritma ini banyak digunakan untuk *feature selection* karena kemampuannya untuk secara efisien menjelajahi ruang pencarian yang besar dan menemukan solusi yang mendekati optimal. Algoritma meta-heuristik dapat digunakan untuk mencari subset fitur yang optimal dengan mengevaluasi kualitas solusi kandidat berdasarkan fungsi kecocokan. Fungsi kecocokan biasanya didasarkan pada metrik kinerja, seperti akurasi, F1-score, atau AUC.

Algoritma meta-heuristik yang digunakan dalam penelitian ini yaitu *evolutionary based* dan *swarm based*. *Evolutionary based* menggunakan algoritma differential evolution (DE) dan flower pollination (FP) algoritma sementara *swarm based* menggunakan algoritma whale optimization (WO) dan grey wolf (GW). DE menggunakan pendekatan berbasis populasi dan mencari solusi optimal dengan memperbaiki solusi kandidat ecara iteratif melalui proses mutasi dan seleksi. Tanabe mengembangkan algoritma DE dengan memperhitungkan histori dari pemilihan parameter dan menghasilkan performa yang baik (Tanabe & Fukunaga, 2013).

FP adalah algoritma yang terinspirasi dari proses penyerbukan bunga pada tumbuhan. Proses pencarian solusi dimulai dengan menginisialisasi populasi awal, kemudian dalam setiap iterasi solusi yang lebih baik dihasilkan dengan menggabungkan sifat-sifat dari solusi-solusi yang ada dalam populasi dengan probabilitas tertentu. Yang telah mengembangkan algoritma FP dan menemukan bahwa FP lebih baik dari genetic algorithm dan particle swarm optimization (Yang, 2012).

WO merupakan algoritma yang terinspirasi dari prilaku berburu paus oleh kelompok paus pembunuh. Mirjalili telah mengembangkan algoritma optimisasi dengan meniru prilaku paus dalam berburu (Mirjalili & Lewis, 2016). GW merupaka algoritma yang meniru pergerakan kelompok serigala dalam berburu. Gupta telah mengembangkan algoritma WO dan menunjukan hasil yang baik dalam permasalahan optimisasi (Gupta & Deep, 2019).

Penelitian ini mengembangkan proses *feature selection* dengan menggunakan algoritma meta-heuristik. Data yang digunakan merupakan data madelon. Proses klasifikasi dilakukan dengan algoeritma RF, kNN dan SVM sementara proses *feature selection* dilakukan dengan algoritma DE, FP, WO dan GW.

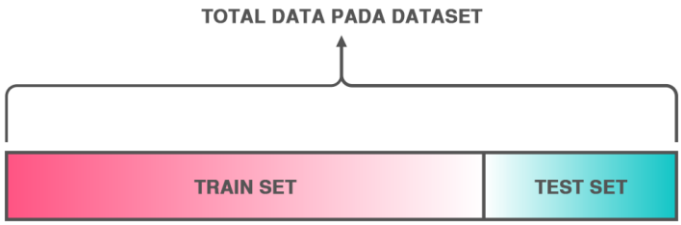
METODE PENELITIAN

Pengumpulan Data

Data yang digunakan adalah data Madelon. Madelon merupakan data buatan yang terdiri dari 500 fitur dan 2 kelas. Data tersebut tersedia pada NIPS 2003 *feature selection challenge* (Guyon, dkk., 2004).

Pre-processing Data

Data terkumpul diproses terlebih dahulu, bertujuan dalam mempermudah *machine* mempelajari pola, data yang ada dibersihkan terlebih dahulu. Dataset yang telah dibersihkan dan diproses kemudian siap kita latih dengan *machine learning*. Untuk mengetahui apakah model *machine learning* kita bagus atau tidak adalah dengan mengujinya pada kasus atau data baru yang belum dikenali oleh model. Pilihan yang lebih baik adalah dengan membagi dataset menjadi 2 bagian yaitu data training dan data testing. Gambar 1 adalah gambaran bagaimana total data pada dataset dibagi menjadi dua bagian: train set dan test set.

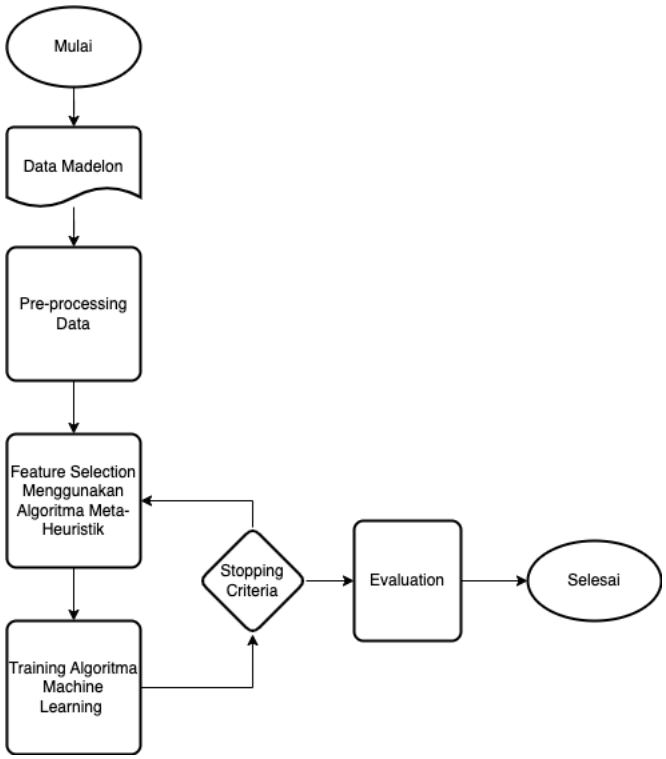


Gambar 1 Pembagian Dataset

Data testing diambil dengan proporsi tertentu. Pada praktiknya, pembagian data training dan data testing yang paling umum adalah 80:20, 70:30, atau 60:40, tergantung dari ukuran atau jumlah data. Namun, untuk dataset berukuran besar, proporsi pembagian 90:10 atau 99:1 juga umum dilakukan.

Membangun Model

Model merupakan perpaduan antara model *machine learning* dan algoritma meta-heuristik. Algoritma meta-heuristik digunakan untuk *feature selection* sementara model *machine learning* digunakan untuk proses klasifikasi. Model lengkap dapat dilihat pada gambar 2.



Gambar 2 Alur Model

HASIL DAN PEMBAHASAN

Pembuatan Model

Penilitian ini algoritma meta-heuristik digunakan untuk memilih fitur yang paling berpengaruh terhadap proses klasifikasi. *Feature selection* bertujuan untuk mendapatkan fitur paling informatif sehingga meningkatkan akurasi dan efesiensi. Hasil *feature selection* berupa subset fitur yang digunakan sebagai input pada model *machine learning*. Algoritma meta-heuristik yang digunakan adalah Differential Evolution, Flower Pollination, Grey Wolf dan Whale Optimization serta Model Machine Learning yang digunakan adalah Random Forest, k-Nearest Neighbors dan Support Vector Machine. Penelitian ini dilakukan beberapa skema antara proses klasifikasi dan *feature selection*. Skema terdiri dari kombinasi antara model *machine learning* dan algoritma meta-heuristik. Berikut skema dari kombinasi model *machine learning* dan algoritma meta-heuristik.

Tabel 1. Skema Model Klasifikasi dan *Feature Selection*

Skema
Random Forest – Differential Evolution (RF-DE)
Random Forest – Flower Pollination (RF-FP)
Random Forest – Grey Wolf (RF-GW)
Random Forest – Whale Optimization (RF-WO)
k-Nearest Neighbors – Differential Evolution (KNN-DE)
k-Nearest Neighbors – Flower Pollination (KNN-FP)
k-Nearest Neighbors – Grey Wolf (KNN-GW)
k-Nearest Neighbors – Whale Optimization (kNN-WO)
Support Vector Machine – Differential Evolution (SVM-DE)
Support Vector Machine – Flower Pollination (SVM-PA)
Support Vector Machine – Grey Wolf (SVM-GW)
Support Vector Machine – Whale Optimization (SVM-WO)

Pengujian

Data hasil pre-processing digunakan pada tahap *training* dan *testing*. Data madelon terdiri dari 500 fitur dan 2600 *record*. Model machine learning bertujuan untuk melakukan binary klasifikasi pada data Madelon, output dari model merupakan angka biner yaitu 0 atau 1. Data hasil pre-processing dibagi menjadi 75% untuk data *training* dan 25% untuk data *testing*. Model klasifikasi dilakukan proses *training* dan *testing* menggunakan data tersebut. Setiap model dibangun dengan parameter seperti pada Tabel 2. Hasil akurasi dari setiap model klasifikasi yang ada terlihat pada Tabel 3.

Tabel 2. Parameter Setiap Model Klasifikasi

Model Klasifikasi	Parameter	Nilai
RF	n_estimator	100
	Minimun samples split	2
	Minimum samples leaf	1
KNN	Number of neighbors	5
	Metric	Minkowski
SVM	Penalty parameter of the error term *(C)*	1
	Degree	3
	Epsilon	0.2
	Gamma	Scale
	Toleransi	0.001
	Kernel	Radial Basic Function (RBF)

Tabel 3. Akurasi Setiap Model Klasifikasi

Model Klasifikasi	Akurasi	Features
RF	0.69	500
KNN	0.71	500
SVM	0.65	500

Tabel 4. Pengaturan Parameter Setiap Algoritma

Algoritma Meta-Heuristik	Parameter	Nilai
DE	Initial Weighting Factor	0.5
	Initial Cross-over Probability	0.5
	Population Size	20
	Epoch	20
FP	Switch Probability	0.8
	Levy Multiplier	0.2
	Population Size	20
	Epoch	20
GW	Population Size	20
	Epoch	20
WO	Maximum iterations of each feedback	10
	Population Size	20
	Epoch	20

Pembahasan

Feature selection dilakukan menggunakan algoritma meta-heuristik. Pengaturan parameter pada setiap algoritma meta-heuristik terlihat pada Tabel 4. Model klasifikasi yaitu KNN, RF dan SVM menjadi fungsi objektif bagi algoritma meta-heuristik. Fungsi objektif dievaluasi pada setiap iterasi, iterasi dibatasi oleh jumlah epoch. Pada proses iterasi algoritma mencari nilai yang optimal dengan pemilihan fitur berdasarkan aturan masing-masing algoritma. Tabel 5 merupakan nilai akurasi dan jumlah fitur yang digunakan pada proses klasifikasi.

Tabel 5 menunjukan bahwa algoritma meta-heuristik berhasil menjalankan tugas sebagai *feature selector*, terlihat dari jumlah fitur yang berkurang. Keberhasilan model dalam melakukan klasifikasi meningkat setelah dilakukan *feature selection*, hal ini menunjukan algoritma meta-heuristik berhasil dalam memilih fitur yang paling informatif pada data. Secara keseluruhan setiap skema menunjukan keberhasilan karena meningkatkan akurasi dan mengurangi jumlah fitur yang digunakan pada proses klasifikasi.

Tabel 5. Akurasi Setiap Skema

	Akurasi	Features
KNN-DE	0.76	116
KNN-FP	0.73	229
KNN-GW	0.72	27
KNN-WO	0.64	247
SVM-DE	0.79	24
SVM-FP	0.71	53
SVM-GW	0.73	143
SVM-WO	0.74	63
RF-DE	0.75	206
RF-FP	0.75	28
RF-GW	0.74	227
RF-WO	0.73	64

Nilai akurasi tiap skema merupakan nilai terbaik yand dicapai selama iterasi. Meta-heuristik berbasis evolusi DE dan FP memberikan nilai akurasi lebih tinggi dibandingkan dengan meta-heuristik berbasis perlaku kawanan GW dan WO. Secara keseluruhan DE merupakan algoritma meta-heuristik yang paling berhasil meningkatkan akurasi dengan fitur yang lebih sedikit. Skema paling berhasil adalah skema SVM-DE ditunjukan dengan nilai akurasi tinggi dengan fitur terendah.

KESIMPULAN

Feature selection dapat meningkatkan akurasi model *machine learning*. Setelah dilakukan *feature selection* model KNN, RF dan SVM menunjukan peningkatan akurasi. Algoritma meta-heuristik dapat dengan baik memilih fitur yang paling berpengaruh, terlihat dari berkurangnya fitur dan bertambahnya akusrai, Skema KNN-DE merupakan skema dengan nilai akurasi tinggi dengan penggunaan fitur terendah. Penelitian ini belum melakukan uji coba untuk parameter berbeda pada setiap algoritma dan melakukan uji coba pada data lain. Peneliatian selanjutnya dapat dilakukan dengan data berbeda dan parameter yang berbeda untuk melihat tingkat keberhasilan *feature selection* menggunakan algoritma meta-heuristik.

DAFTAR PUSTAKA

Cervantes, J., Garcia-Lamont, F., Rodriguez-Mazahua, L., dan Lopez, A. (2020). A Comprehensive Survey on Supprot Vector Machine Clasification: Applications, Chalenges and Trends. Neurocomputing, (189-215). https://doi.org/10.1016/j.neucom.2019.10.118.

Diaz-Uriarte, R., dan de Andres, R. A. (2006). Gene Selection and Classification of Microarray Data Using Random Forest. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-7-3.

Gupta, S., & Deep, K. (2019). A Novel Random Walk Grey Wolf Optimizer. Swarm and evolutionary computation, 44, 101-112. https://doi.org/10.1016/j.swevo.2018.01.001.

Mirjalili, S., & Lewis, A. (2016). The Whale Optimization Algorithm. Advances in Engineering Software, (51-67). https://doi.org/10.1016/j.advengsoft.2016.01.008.

Nababan, A. A., Sitompul, O. S., dan Tulus. (2018). Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio. Journal of Physics: Conference Series.

Yang, X. S. (2012). Flower Pollination Algorithm for Global Optimization. International conference on unconventional computing and natural computation (240-249). https://doi.org/10.1007/978-3-642-32894-7_27.

Tanabe, R & Fukunaga, A. (2013). Success-History Based Parameter Adaptation for Differential Evolution. IEEE Congress on Evolutionary Computation. https://doi.org/10.1109/CEC.2013.6557555.