Min-Max Scaling, Z-Score, Decimal Scaling

Penelitian tentang pengaruh teknik normalisasi terhadap algoritma KNN telah banyak dilakukan untuk mengeksplorasi cara meningkatkan kinerja klasifikasi. Salah satu penelitian penting adalah studi oleh Alshdaifat et al. (2021) yang mengevaluasi dampak berbagai teknik normalisasi, termasuk Min-Max Normalization, Z-Score Normalization, dan Decimal Scaling Normalization, terhadap kinerja algoritma klasifikasi seperti Support Vector Machines (SVM) dan Artificial Neural Networks (ANN). Meskipun penelitian ini berfokus pada SVM dan ANN, temuan yang menunjukkan pengaruh signifikan teknik normalisasi terhadap performa algoritma klasifikasi juga relevan untuk algoritma K-Nearest Neighbor (KNN), mengingat KNN sangat sensitif terhadap skala fitur dalam perhitungan jaraknya.

Penelitian ini dilakukan pada 18 dataset benchmark dari berbagai domain, termasuk kesehatan, bisnis, dan fisika. Hasilnya menunjukkan bahwa Z-Score Normalization sering kali memberikan hasil terbaik, khususnya karena kemampuannya menangani outlier, sedangkan Decimal Scaling Normalization dianggap kurang efektif dalam meningkatkan akurasi model secara keseluruhan. Temuan ini menjadi dasar penting untuk mengeksplorasi pengaruh normalisasi pada algoritma KNN.

SITASI: The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms’ Performance

Min-Max Scaling, Z-Score, Decimal Scaling

Penelitian lainnya dilakukan oleh **Muasir Pagan et al.** pada tahun 2023, yang menyelidiki dampak teknik skala data pada kinerja algoritma K-Nearest Neighbor (k-NN) menggunakan sepuluh dataset dari berbagai domain. Tiga teknik skala yang dievaluasi dalam penelitian ini adalah **min-max normalization**, **Z-score**, dan **decimal scaling**. Hasil penelitian menunjukkan bahwa pilihan teknik skala memiliki dampak signifikan terhadap kinerja algoritma k-NN, dengan Z-score dan decimal scaling umumnya memberikan hasil yang lebih baik dibandingkan min-max normalization.

Secara khusus, penelitian ini menemukan bahwa Z-score secara konsisten memberikan akurasi, presisi, recall, dan F1-score yang lebih tinggi di sebagian besar dataset yang diuji. Meskipun min-max normalization menunjukkan performa yang baik pada beberapa dataset tertentu, seperti dataset Airfoil self-noise dan Concrete compressive strength, Z-score dan decimal scaling lebih unggul dalam konteks yang lebih luas. Temuan ini menekankan pentingnya pemilihan teknik skala yang tepat berdasarkan karakteristik spesifik dari dataset, memberikan wawasan berharga bagi praktisi dan peneliti di bidang pembelajaran mesin.

SITASI: Investigating the impact of data scaling on the KNN algorithm

Min-Max Scaling, Z-Score, Robust Scaler

Penelitian lainnya dilakukan oleh **Md Manjurul Ahsan et al.** pada tahun **2021**, yang mengeksplorasi pengaruh metode skala data terhadap algoritma pembelajaran mesin dan kinerja model. Penelitian ini menganalisis beberapa teknik skala, termasuk **min-max normalization**, **Z-score normalization**, dan **robust scaling**, dengan tujuan untuk memahami bagaimana masing-masing metode mempengaruhi akurasi dan efisiensi berbagai algoritma pembelajaran mesin, seperti k-NN, Decision Tree, dan Support Vector Machine (SVM). Dataset yang digunakan mencakup berbagai domain, memberikan gambaran yang komprehensif tentang dampak teknik skala pada kinerja model.

Hasil penelitian menunjukkan bahwa metode skala data memiliki pengaruh signifikan terhadap kinerja algoritma pembelajaran mesin. Secara umum, Z-score normalization dan robust scaling menunjukkan hasil yang lebih baik dibandingkan dengan min-max normalization, terutama pada dataset yang memiliki outlier. Penelitian ini juga menemukan bahwa pemilihan teknik skala yang tepat dapat meningkatkan akurasi model secara signifikan, serta mengurangi waktu pelatihan. Temuan ini memberikan wawasan penting bagi praktisi dan peneliti dalam memilih metode skala yang sesuai untuk meningkatkan kinerja model pembelajaran mesin mereka.

SITASI: Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance

Min-Max Scaling

Penelitian lainnya dilakukan oleh Peshawa J. Muhammad Ali (2022), yang mengevaluasi dampak Min-Max Normalization terhadap kinerja algoritma K-Nearest Neighbor (KNN) dalam tugas regresi menggunakan delapan ukuran kesamaan, yaitu City Block, Euclidean, Chebychev, Cosine, Correlation, Hamming, Jaccard, dan Mahalanobis. Penelitian ini dilakukan pada lima dataset benchmark dari UCI Machine Learning Repository yang mencakup berbagai domain, seperti fisika, ilmu hayat, energi, teknik sipil, dan bisnis. Hasil penelitian menunjukkan bahwa Min-Max Normalization memiliki pengaruh signifikan pada ukuran kesamaan berbasis jarak, seperti City Block, Euclidean, dan Chebychev, terutama pada dataset dengan perbedaan besar dalam rentang fitur. Namun, teknik ini tidak memengaruhi ukuran kesamaan seperti Hamming, Jaccard, dan Mahalanobis, karena sifat matematis mereka tidak bergantung pada skala fitur. Selain itu, ukuran kesamaan Mahalanobis secara konsisten memberikan hasil terbaik di semua dataset, baik sebelum maupun setelah normalisasi.

SITASI: Investigating the Impact of Min-Max Data Normalization on the Regression Performance of K-Nearest Neighbor with Different Similarity Measurements

Z-Score

Penelitian lainnya dilakukan oleh **Umarani P** dan **Viswanathan P** pada tahun 2022, yang berfokus pada penerapan **normalisasi Z-Score** dalam sistem diagnosis otomatis kanker darah menggunakan algoritma K-Nearest Neighbor (k-NN). Penelitian ini dilakukan pada dataset leukemia yang terdiri dari 22 fitur statistik dan geometris yang diambil dari sel darah putih (WBC), nukleus, dan sitoplasma. Fitur-fitur tersebut mencakup atribut seperti area WBC, perimeter, eksentrisitas, dan lainnya, yang diekstraksi dari gambar apusan darah. Hasil penelitian menunjukkan bahwa penerapan normalisasi Z-Score dapat meningkatkan akurasi sistem diagnosis otomatis.

Dalam penelitian ini, peneliti mengevaluasi kinerja algoritma k-NN dengan menggunakan ukuran jarak maksimum, yaitu Chebyshev. Akurasi tertinggi yang dicapai adalah **97,92%** ketika k=1. Namun, setelah melakukan tuning parameter melalui metode grid search, akurasi sistem meningkat menjadi **98,65%**. Temuan ini menunjukkan bahwa ukuran jarak Chebyshev memberikan kinerja terbaik dibandingkan dengan ukuran jarak lainnya seperti Euclidean, Minkowski, dan Manhattan. Penelitian ini menegaskan pentingnya teknik normalisasi dan pemilihan ukuran jarak yang tepat dalam meningkatkan akurasi sistem diagnosis otomatis untuk kanker darah, terutama pada dataset yang memiliki variasi besar dalam fitur.

SITASI: Z-Score Normalized Features with Maximum Distance Measure Based k-NN Automated Blood Cancer Diagnosis System

Log Transformation

Penelitian lainnya dilakukan oleh **Shuangbin Xu et al.** pada tahun **2021**, yang mengeksplorasi pengaruh **log transformation** dalam teknik normalisasi data. Penelitian ini menunjukkan bahwa log transformation adalah metode yang umum digunakan untuk mengurangi skewness dalam distribusi data, terutama dalam konteks data yang memiliki outlier. Meskipun log transformation dapat membantu dalam mengubah distribusi data menjadi lebih mendekati distribusi normal, penelitian ini juga mencatat bahwa data yang telah ditransformasi tidak berada pada skala yang sama dengan data asli, yang dapat mempengaruhi interpretasi hasil.

Hasil penelitian menunjukkan bahwa log transformation dapat mengurangi efek dari outlier, tetapi juga dapat menghilangkan informasi penting yang terkandung dalam data asli. Oleh karena itu, penting untuk mempertimbangkan jenis data yang digunakan sebelum menerapkan log transformation. Penelitian ini juga menyoroti bahwa meskipun log transformation dapat meningkatkan akurasi model dalam beberapa kasus, pemilihan teknik normalisasi yang tepat harus disesuaikan dengan karakteristik spesifik dari dataset yang dianalisis. Temuan ini memberikan wawasan penting bagi peneliti dan praktisi dalam memilih metode normalisasi yang sesuai untuk meningkatkan kinerja model analisis data mereka.

SITASI: Use *ggbreak* to Effectively Utilize Plotting Space to Deal With Large Datasets and Outliers