Penerapan Algoritma EDISA (Extended Dimension Iterative Signature Algorithm) Triclustering pada Data Ekspresi Gen Tiga Dimensi = Application of EDISA (Extended Dimension Iterative Signature Algorithm) Triclustering Algorithm in Three-Dimensional Gene Expression Data

Dwi Aji Apriana, author

Abstrak

Triclustering merupakan salah satu metode data mining yang juga merupakan pengembangan dari metode biclustering dan clustering. Metode tersebut mengelompokkan set data berupa matriks tiga dimensi (gen, kondisi, dan waktu) menjadi kelompok-kelompok submatriks yang memiliki kesamaan satu sama lain. Salah satu algoritma dari analisis triclustering adalah Extended Dimension Iterative Signature Algorithm (EDISA). Algoritma ini mempertimbangkan jarak Pearson antara tiap gen dan kondisi terhadap vektor rata-rata sebagai ukuran kemiripan. Proses pertama dari EDISA adalah langkah preprocessing yaitu menghapus gen yang memiliki nilai ekspresi gen yang berbeda sangat signifikan dengan nilai ekspresi gen lainnya. Lalu langkah selanjutnya yaitu memilih sebanyak s sampel gen dengan cara memilih satu gen secara random untuk menjadi seed gen, lalu mencari sebanyak s-1 gen yang memiliki jarak Pearson terdekat dengan seed gen tersebut. Tahap berikutnya membuat vektor bobot gen dan kondisi, lalu memasangkannya dengan sampel gen yang telah terpilih, kemudian menghitung vektor rata-ratanya. Proses selanjutnya yaitu proses iterasi di mana setiap gen dan kondisi yang memiliki jarak Pearson terhadap vektor rata-rata di atas ambang batas tertentu (TG dan TG, keduanya merupakan ukuran seberapa baik keselarasan suatu gen dan kondisi terhadap rata-rata kandidat tricluster) harus dihapus karena dianggap tidak memiliki kemiripan yang cukup dengan anggota tricluster lain pada setiap iterasinya. Proses selanjutnya adalah postprocessing yang bertujuan untuk menggabungkan tricluster yang memiliki kemiripan untuk dijadikan tricluster yang lebih besar dan dijadikan sebagai kumpulan tricluster final. Algoritma ini diterapkan pada data ekspresi gen penyakit paru-paru. Penerapan algoritma tersebut menggunakan beberapa skenario dengan nilai Tg dan TG yang berbeda. Hasil dari penerapan pada data ekspresi gen penyakit paru-paru diperoleh bahwa semakin besar nilai TG, maka jumlah gen yang dapat masuk ke dalam tricluster makin banyak, dan semakin besar nilai TG, maka jumlah kondisi yang dapat masuk ke dalam tricluster juga makin banyak. Selain itu, dilakukan evaluasi dari tricluster menggunakan nilai Tricluster Diffusion Score (TD Score) untuk mencari skenario terbaik. Didapat bahwa skenario terbaik merupakan skenario dengan nilai Tg=0,3 dan nilai TG=0,2. Melalui algoritma ini dapat dideteksi gen-gen yang dapat membedakan karakteristik pasien yang berpenyakit paru-paru dan pasien yang sehat.