INTISARI

Optimisasi K-Means Data Nonlinear Separable Dengan Kernel Principal Component Analysis

Cori Pitoy
10/307054/SPA/00339

Klasterisasi K-Means merupakan salah satu metode klasterisasi *non-hierarchical*, yang berupaya menemukan *k* klaster sedemikian sehingga data dengan karakteristik yang sama dikelompokkan ke dalam klaster yang sama, sedangkan data dengan karakteristik berbeda dikelompokkan ke dalam klaster lainnya dengan tujuan meminimalkan varians data dalam klaster dan memaksimalkan varians data antar klaster. K-Means akan bekerja dengan baik hanya untuk data yang berbentuk bulat, terpisahkan secara linear (*linearly separable*) dan klaster yang memiliki ukuran/densitas data yang sama.

Permasalahannya, tidak semua data memiliki struktur yang berbentuk bulat dan dapat terpisahkan secara linear sehingga dapat dengan mudah dilakukan klasterisasi seperti pada prosedur klasterisasi K-Means. Penelitian ini dilakukan melalui *preprocessing data*, pemilihan banyaknya PC yang digunakan dan pemilihan besaran parameter bandwidth kernel Gaussian (γ) yang sesuai pada pemetaan KPCA ke ruang fitur, sehingga akurasi hasil klasterisasi K-Means, baik validitas internal maupun eksternal menjadi optimum dan konsisten. Hasil penelitian ini diharapkan dapat mengatasi beberapa kelemahan metode klasterisasi sebelumnya, yang hanya unggul untuk data terpisah secara linear dan data berbentuk bulat seperti pada Info K-Means dari Wu (2012) dan klaster dengan ukuran/densitas sama pada K-Means (MacQueen, 1967).

Prosedur klasterisasi data *nonlinearly separable* dimodifikasi menjadi dua tahap. Tahap pertama, data $x_1, \dots, x_n$ pada di ruang input $\mathbb{R}^d$ diekstrak melalui pemetaan menggunakan fungsi kernel $\Phi(z_i)$ ke ruang fitur $\mathcal{F}$ (*feature space*) yang berdimensi lebih tinggi dengan menggunakan kernel Gaussian. Pada ruang fitur, dilakukan data ekstraksi dan reduksi dimensi menggunakan PCA, kemudian dilakukan perhitungan klasterisasi K-Means. Untuk meningkatkan akurasi hasil klasterisasi, terlebih dahulu dilakukan standarisasi terhadap data di ruang input, mencari ukuran bandwidth kernel Gaussian (γ) pada pemetaan KPCA dan pemilihan banyaknya PC yang digunakan pada klasterisasi K-Means ruang fitur. Ukuran bandwidth kernel Gaussian (γ) dipilih berdasarkan hasil interpolasi polinomial menggunakan beberapa titik dalam rentang 0.1 s/d 1.0, sedangkan pemilihan banyaknya PC(p) dilakukan melalui perbandingan grafis, yaitu PC dengan hasil yang mengoptimumkan SSW(%) dan Entropy secara bersama-sama.

Prosedur klasterisasi K-Means terhadap data Iris dilakukan sebanyak n=100 menggunakan ukuran bandwidth kernel Gaussian γ = 0.24 pada pemetaan KPCA, memberikan hasil klasterisasi K-Means yang optimal dan konsisten. Validitas eksternal hasil klasterisasi K-Means diperoleh dengan rata-rata nilai

entropy $\bar{x}_{\gamma = 0.24}$ = 0.05 dan varians $s^2_{\gamma = 0.24}$= 0. Validitas internal diperoleh dengan nilai rata-rata SSW (%) $\bar{x}_{\gamma = 0.24}$= 11.9 dan varians $s^2_{\gamma = 0.24}$= 0. Hasil yang sama diperoleh untuk n = 200 dan n = 250.

Hasil ini jauh lebih baik dan konsisten dibandingkan dengan hasil klasterisasi dengan algoritma K-Means (MacQueen, 1967), baik menggunakan dataset original maupun terstandar. Entropy hasil klasterisasi K-Means menggunakan data original menghasilkan nilai rata-rata $\bar{x}_{\mathrm{o}}$ = 0.127 dan varians s^2_{o}= 0.034. Entropy dari data terstandar menghasilkan nilai rata-rata $\bar{x}_{\mathrm{s}}$ = 0.312 dan varians s^2_{s}= 0.005. SSW (%) hasil klasterisasi K-Means menggunakan data original menghasilkan nilai rata-rata $\bar{x}_{\mathrm{o}}$ = 13.532 dan varians s^2_{o}=19.476 dan untuk data terstandar menghasilkan nilai rata-rata $\bar{x}_{\mathrm{s}}$ = 24.135 dan varians s^2_{s}= 13.167.

Kata Kunci : *K-Means, KPCA, Bandwidth, Validitas, Entropy, SSW, Nonlinearly Separable, Input, Fiture.*

ABSRACT

K-Means Optimization of Nonlinear Separable Data using Kernel of Principal Component Analysis

Cori Pitoy
10/307504/SPA/0039

K-Means Clustering is one of non-hierarchical clustering methods aiming at finding *k* clusters such that data which have similar characteristics are group into the same clusters whereas data which are of not similar characteristics are group on different clusters by minimalizing variance in clusters and maximizing variance between clusters. K-Means works well when data are of round form, linearly separable and clusters have equal size or density.

The problem is that not all data have round structure and are separated linearly such that it is easy to do clustering as in K-Means procedure. This research is done by prepocessing data, choosing number of PC (principal components) used and choosing Gaussian Kernel bandwith parameter (γ) which suits in mapping KPCA into feature space, such that accuracy of the K-means result, internal and eksternal validity, are optimum and consistent. The result of this research is expected to overcome some obstacles on previous clustering methods which are superior for only linearly separable data and in round form such as in K-Means (Wu, 2012) and clusters with equal size or density on K-Means (MacQueen,1967).

Clustering method of nonlinearly separable data is modified into two stages. Stage one, data $x_{1,\cdots,}x_m$ in input space $\mathbb{R}^d$ is extracted through mapping by kernel function $\Phi(z_i)$ into feature space $\mathcal{F}$ which has higher dimension by Gaussian Kernel. In the feature space, data extraction dan dimension reduction are applied by PCA, then K-Means clustering is applied. To increase the clustering accuracy, standardizing data is done first in input space, determine the size of Gaussian Kernel bandwith parameter (γ) on KPCA mapping and choose the number of PC which will be used later in K-Means clustering in feature space. The size of Gaussian Kernel bandwith parameter (γ) is chosen based on polinomial interpolation using some points ranging from 0.1 up to 1.0, whereas number of PC (*p*) is done by comparing graphics, that is PC with results that optimize SSW (%) and Entropy together.

K-Means clustering procedure on Iris data applied n=100 times using Gaussian Kernel bandwith with size γ = 0.24 on KPCA mapping, gives optimal and consistent K-Means clustering results. External validity of K-Means clustering obtained are Entrophy mean $\bar{x}_{\gamma=0.24} = 0.05$ and variance $s_{\gamma=0.24} = 0.000$. Internal validity obtained are SSW (%) mean $\bar{x}_{\gamma=0.24} = 11.9$ and variance $s_{\gamma=0.24} = 0.000$. Similar results are found for n=200 and n=250.

The results are far better and consistent compared with K-Means Clustering algorithm (MacQueen, 1967), for original dataset and even for standardized data set. The entrophy of K-Means Clustering applied on original data has mean and variance, $\bar{x}_0 = 0.127$ and $s_0^2 = 0.034$, respectively. The entrophy of standardized data has mean and variance, $\bar{x}_s = 0.312$ $s_0^2 = 0.034$, respectively. Meanwhile, the SSW (%) of K-Means clustering on original data has mean and variance, $\bar{x}_0 = 13.532$ and $s_0^2 = 19.476$, respectively, and for standardized data, these are $\bar{x}_s = 24.135$ $s_0^2 = 13.167$, respectively.

Key Words : *K-Means, KPCA, Bandwidth, Validity, Entropy, SSW, Nonlinearly Separable, Input, Fiture*