Indonesian Journal of Technology, Informatics and Science (IJTIS) DOI: 10.24176/ijtis.v1i2.4939
Vol. 1, No. 2, Juni 2020, hlm. 55-58 p-ISSN: 2715-940X
e-ISSN: 2721-4303

PENERAPAN *PRINCIPAL COMPONENT ANALYSIS* UNTUK PENINGKATAN KINERJA ALGORITMA *DECISION TREE* PADA *IRIS DATASET*

Putri Kurnia Handayani[1]

[1]Fakultas Teknik Universitas Muria Kudus
Email: [1]putri.kurnia@umk.ac.id

(Naskah masuk: 24 Juni 2020, diterima untuk diterbitkan: 29 Juni 2020)

Abstrak

Data mining merupakan salah bidang ilmu yang bermanfaat untuk pengenalan pola/*knowledge* yang tersimpan dalam database. Klasifikasi merupakan salah satu peran dalam bidang data mining. Termasuk ke dalam *supervised learning*, klasifikasi digunakan untuk memprediksi objek yang belum memiliki kelas/label. Penggunaan algoritma *decision tree* untuk proses mining dataset bunga iris dikarenakan kemudahan dalam representasi *knowledge* yang dihasilkan. Selain itu, *decision tree* juga termasuk ke dalam *eager learner* sehingga akurasi dari *knowledge* yang dihasilkan lebih baik. Penggunaan *principal component analysis* (PCA) dalam optimasi algoritma *decision tree*, dilakukan saat *preprocessing* dataset. PCA berfungsi untuk mereduksi dimensi, fitur yang saling berkorelasi akan dipertahankan. Penggunaan dataset publik bunga iris diambil dari UCI Repository. Berdasarkan hasil perhitungan, akurasi algoritma decision tree setelah dilakukan optimasi dengan PCA terhadap dataset bunga iris sebesar 95.33%.

Kata kunci: *data mining, klasifikasi, decision tree, PCA*

IMPLEMENTATION OF PRINCIPAL COMPONENT ANALYSIS FOR IMPROVING DECISION TREE ALGORITHM PERFORMANCE IN IRIS DATASET

Abstract

Data mining is a useful field of knowledge for pattern recognition / knowledge stored in databases. Classification is one of the roles in the field of data mining. Included in supervised learning, classification is used to predict objects that do not have a class / label. The use of decision tree algorithms for the iris dataset mining process is due to the ease in representing the knowledge generated. In addition, the decision tree is also included in the eager learner so that the accuracy of the knowledge generated is better. The use of principal component analysis (PCA) in the decision tree algorithm optimization is done during the preprocessing dataset. PCA functions to reduce the dimensions, correlated features will be maintained. The use of the iris public dataset is taken from the UCI Repository. Based on the calculation results, the accuracy of the decision tree algorithm after optimization with the PCA of the iris dataset is 95.33%

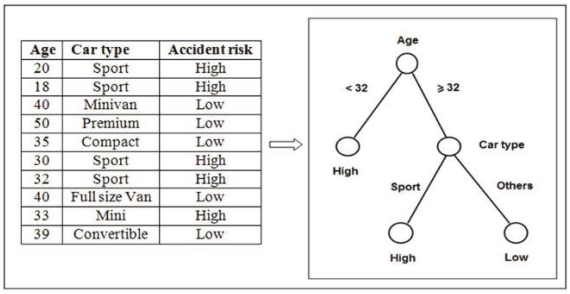
Keywords: *data mining, classification, decision tree, PCA*

1. PENDAHULUAN

Data mining didefinisikan sebagai pencarian pola secara otomatis dalam database yang besar, menggunakan teknik komputasi dari statistik, *machine learning* dan *pattern recognition* (Gorunescu, 2011). Data mining adalah tentang bagaimana menyelesaikan masalah dengan cara menganalisis data yang tersedia dalam database ((Witten, et al., 2011)

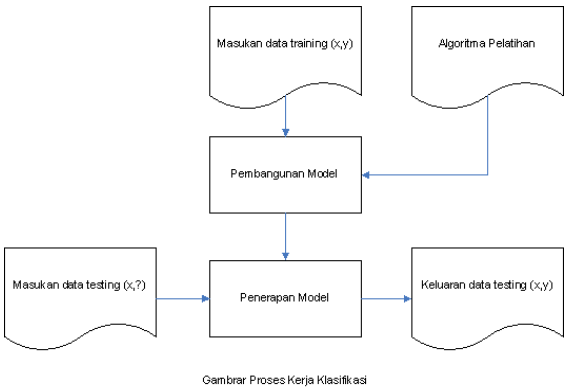
Klasifikasi merupakan salah satu dari lima peran utama data mining. Termasuk ke dalam kategori *supervised learning*, klasifikasi merupakan proses prediksi terhadap objek yang belum mempunyai kelas/label. *Decision tree* (DT) adalah *classifier* yang dinyatakan sebagai partisi rekursif dari dataset ((Maimon & Rokarch, 2010). *Decision tree* terdiri dari akar, cabang dan daun. *Decision tree* populer digunakan untuk proses klasifikasi karena kemudahan dalam representasi pola dataset (gambar 1). Selain digambarkan dalam bentuk pohon, hasil dari klasifikasi juga bisa direpresentasikan dalam bentuk aturan (*rules*). Pola yang terbentuk dari proses klasifikasi dapat digunakan sebagai saran bagi

manajemen dalam menentukan aturan atau kebijakan baru.

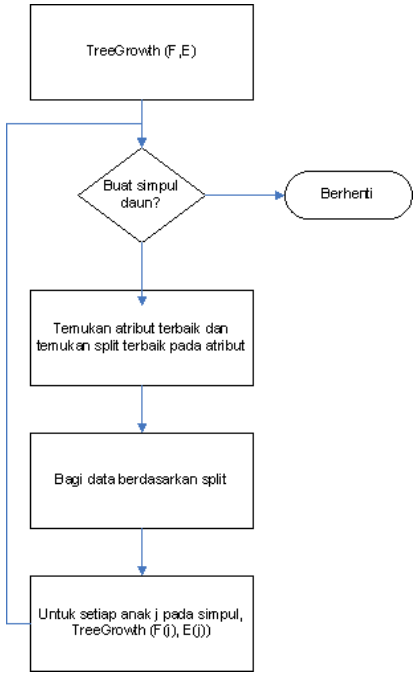


Gambar 1 Skema Decision Tree

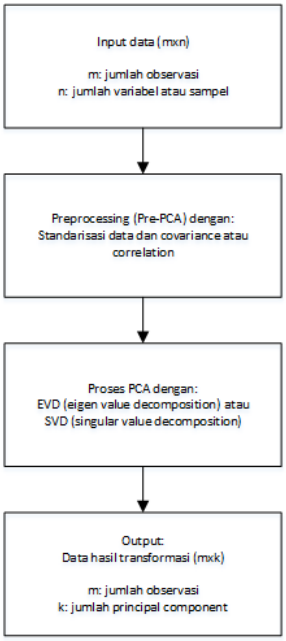
2. METODE PENELITIAN



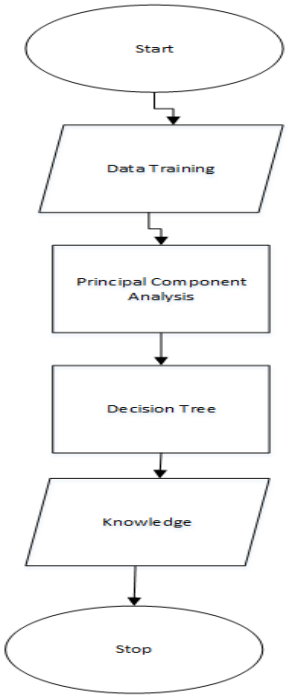
Gambar 2 Alur kerja klasifikasi



Gambar 3 Algoritma Induksi Decision Tree



Gambar 4 Algoritma PCA



Gambar 5 Optimasi Decision Tree dengan PCA

3. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data

Iris dataset merupakan data public yang diambil dari UCI Repository. Data bunga iris terdiri dari 5 atribut, yaitu *sepal length* (a1), *sepal width* (a2), *petal length* (a3), *petal width* (a4) dan label yang berfungsi sebagai kelas/label. Dataset training terdiri dari 150 objek/record yang sudah berada pada kelas/label masing-masing, meliputi kelas iris setosa, iris versicolor dan iris virginica. Sampel dataset bunga iris dapat dilihat pada tabel 1.

Tabel 1 Data sampel bunga Iris

Label	a1	a2	a3	a4
Iris-setosa	5.100	3.500	1.400	0.200
Iris-setosa	4.900	3	1.400	0.200
Iris-setosa	5.400	3.900	1.700	0.400
Iris-setosa	5	3.400	1.500	0.200
Iris-setosa	4.800	3	1.400	0.100
Iris-versicolor	5	2	3.500	1
Iris-versicolor	6.700	3.100	4.400	1.400
Iris-versicolor	5.500	2.400	3.800	1.200
Iris-versicolor	6.200	2.900	4.300	1.300
Iris-versicolor	5.500	2.600	4.400	1.200
Iris-virginica	6.300	3.300	6	2.500
Iris-virginica	5.800	2.700	5.100	1.900
Iris-virginica	6.500	3	5.500	1.800
Iris-virginica	7.400	2.800	6.100	1.900

3.2. Perhitungan Principal Component Analysis (PCA)

Principal Component Analysis (PCA) dapat digunakan untuk mereduksi dimensi suatu data. PCA memberikan hasil yang baik ketika diterapkan pada atribut yang berkorelasi ((Nasution, 2019).

Tahap pertama proses PCA adalah input data. Data disiapkan dalam bentuk matriks ukuran mxn, dimana jumlah variabel n akan berkurang menjadi k jumlah *principal component* yang dipertahankan. Misalkan data yang akan digunakan adalah sebagai berikut (tabel 2).

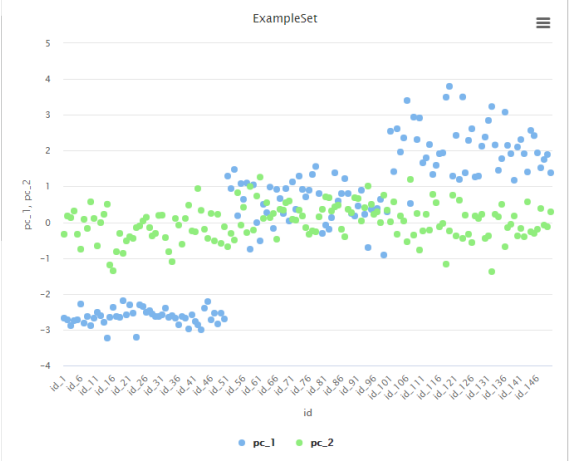
Tabel 2 Sampel data input

a1	a2	a3	a4
5.100	3.500	1.400	0.200
4.900	3	1.400	0.200
5.400	3.900	1.700	0.400
4.800	3	1.400	0.100
5	2	3.500	1
6.700	3.100	4.400	1.400
5.500	2.600	4.400	1.200
6.300	3.300	6	2.500
5.800	2.700	5.100	1.900
6.500	3	5.500	1.800
7.400	2.800	6.100	1.900
5	2.300	3.300	1
5.600	2.800	4.900	2
5.900	3	5.100	1.800

Tahap selanjutnya adalah pre-PCA yaitu dengan melakukan standarisasi dan covariance (tabel 3), dilanjutkan dengan proses PCA. Hasilnya dapat dilihat pada gambar 6. Disini digunakan 2 principal component (pc) yang dipertahankan, yaitu pc1 dan pc2.

Tabel 3 Standarisasi dan covariance

Component	Standard deviation	Proportion of variance
PC1	2.065	0.925
PC2	0.492	0.053
PC3	0.280	0.017
PC4	0.154	0.005



Gambar 6 Hasil PCA

3.3. Optimasi Decision Tree dengan PCA

Decision tree merupakan salah satu algoritma yang populer untuk proses klasifikasi dalam mining. Mengikuti algoritma (gambar 5), proses yang dilakukan setelah mereduksi dimensi dengan PCA adalah menerapkan algoritma d*ecision tree*. Prinsip kerja algoritma *decision tree* adalah menggolongkan objek ke dalam kelas/label yang sudah tersedia. Termasuk ke dalam *supervised learning*, himpunan data bunga iris terdiri dari 3 kelas/label yaitu iris setosa, iris versicolor dan iris virginica. Dataset awal terdiri dari 4 atribut (a1, a2, a3, a4), setelah dilakukan reduksi dimensi dengan PCA hanya 2 atribut yang dipertahankan yakni PC1 dan PC2 (tabel 4).

Tabel 4 Sampel data reduksi dimensi

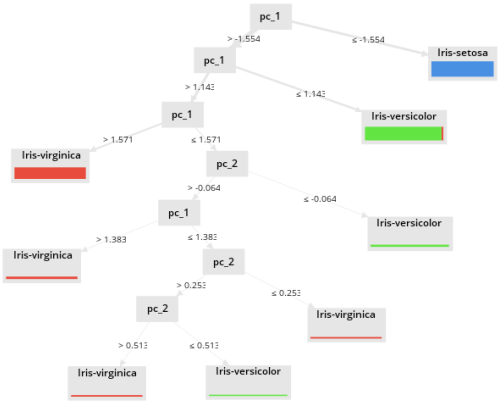
Label	PC1	PC2
Iris-setosa	-2.7	0.2
Iris-setosa	-2.5	-0.5
Iris-setosa	-2.8	0.2
Iris-setosa	-2.5	-0.6
Iris-setosa	-2.7	-0.1
Iris-versicolor	1.3	-0.7
Iris-versicolor	0.9	-0.3
Iris-versicolor	1.5	-0.5
Iris-versicolor	0.2	0.8
Iris-versicolor	1.1	-0.1
Iris-versicolor	0.6	0.4
Iris-virginica	1.3	0.8
Iris-virginica	1.6	0.5
Iris-virginica	1.9	-0.1
Iris-virginica	1.9	0.0
Iris-virginica	3.5	-1.2
Iris-virginica	3.8	-0.3
Iris-virginica	1.3	0.8
Iris-virginica	2.4	-0.4

Berdasarkan reduksi dimensi (tabel 4), langkah selanjutnya adalah pembentukan *tree*. Tahap awal adalah penentuan simpul melalui derajat *impurity*. Derajat *impurity* merupakan ukuran kehomogenan suatu simpul, dimana sebuah simpul dengan derajat impurity tinggi menunjukkan simpul tersebut tidak homogen, begitu sebaliknya. Pengukuran derajat *impurity* dataset menggunakan perhitungan *gini indeks* dengan formula:

$$gini\ (t) = 1 - \sum_j [p(j|t)]^2 \qquad (1)$$

Setelah dilakukan perhitungan nilai *gini indeks* terhadap semua atribut, langkah selanjutnya adalah menghitung nilai gain (Δ) untuk menentukan *best splitting* dengan formula:

$$\Delta = I\,(parent) - \sum_{j=1}^{k} \frac{N(vj)}{N} \mathrm{I(vj)} \qquad (2)$$



Gambar 7 Hasil klasifikasi dengan PCA dan DT

Hasil klasifikasi dengan optimasi algoritma *decision tree* menggunakan PCA berupa *knowledge* atau pola yang direpresentasikan dalam bentuk *tree* (gambar 7). Setelah knowledge atau pola yang ada dalam dataset berhasil diidentifikasi, selanjutnya adalah mengukur akurasi dari algoritma yang digunakan dalam proses data mining. Mengukur akurasi sebuah algoritma dalam klasifikasi dapat menggunakan rumus:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

accuracy: 95.33% +/- 4.50% (micro average: 95.33%)

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	50	0	0	100.00%
pred. Iris-versicolor	0	46	3	93.88%
pred. Iris-virginica	0	4	47	92.16%
class recall	100.00%	92.00%	94.00%	

Gambar 8 Akurasi DT+PCA

Menggunakan tools Rapidminer untuk menggali semua data, akurasi dari algoritma *decision tree* setelah dioptimasi dengan dengan PCA pada proses *preprocessing* (*dimentional reduction*) adalah sebesar 95.33% (gambar 8).

4. KESIMPULAN

Algoritma *decision tree* merupakan salah satu algoritma yang bersifat *eager learner*, dimana algoritma tersebut membaca semua data training yang digunakan untuk kemudian dikenali pola/*knowledge* yang tersimpan dalam data tersebut. Pola/*knowledge* yang sudah dikenali diterapkan pada data testing. *Principal component analysis* (PCA) merupakan metode yang digunakan untuk mereduksi dimensi, yang dilakukan pada tahap *preprocessing*. Akurasi dari algoritma *decision tree* setelah dilakukan optimasi dengan PCA adalah sebesar 95.33% atau meningkat 2.31% dari akurasi *decision tree* murni.

DAFTAR PUSTAKA

Gorunescu, F., 2011. Data Mining Concepts, Models and Techniques. Australia: Springer.

Han, J. & Micheline, K., 2000. Data Mining: Concept and Techniques. Simon Fraser University: Morgan Kaufmann.

Maimon, O. & Rokarch, L., 2010. Data Mining and Knowledge Discovery Handbook. 2 ed. London: Springer.

Nasution, M. Z., 2019. Penerapan Principal Component Analysis (PCA) dalam Penentuan Faktor Dominan yang Mempengaruhi Prestasi Belajar Siswa (Studi Kasus: SMK Raksana 2 Medan). Jurnal Teknologi Informasi, Volume 3, pp. 41-48.

Witten, I. H., Frank, E. & Hall, M. A., 2011. Data Mining Practical Machine Learning Tools and Techniques. 3 ed. Burlington: Elsevier.