

## ANALISIS PENGGUNAAN INDEPENDENT COMPONENT ANALYSIS (ICA) UNTUK REDUKSI DATA BERDIMENSI TINGGI

Mira Dian Ekawati<sup>1</sup>, Adiwijawa<sup>2</sup>, Angelina Prima Kurniati<sup>3</sup>

<sup>1</sup>Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

---

### Abstrak

Data mining merupakan suatu teknik yang dapat memecahkan permasalahan gudang data, sehingga tumpukan data yang tidak memiliki informasi apapun dapat digali dan dianalisa. Tumpukan data tidak dapat langsung diproses, data yang berdimensi tinggi harus dipreprocessing sehingga saat mining data benar-benar memberikan hasil yang maksimal. ICA (Independent Component Analysis) merupakan suatu teknik reduksi pada data numerik yang dapat mengurangi dimensi, dimensi yang baru disebut Independent Component (IC). ICA mereduksi dengan mencari dimensi yang saling independent untuk menjadi dimensi baru. Kondisi ini lebih kuat jika dibandingkan mencari ketidakterhubungan antar dimensi. Fenomena Curse of Dimensionality dapat diselesaikan dengan pengurangan dimensi dengan ICA. Hasil pengujian dengan ICA dapat mereduksi data colon tumor dengan 2000 dimensi dapat diringkas menjadi 60 IC dan data set DLBCL dengan 4026 dimensi dapat diringkas menjadi 46 IC. Kedua dataset memiliki hasil evaluasi clustering terbaik pada 1 IC yaitu dengan nilai SSE sebesar 0.00 dan nilai SSE 0.97. ICA memberikan hasil evaluasi clustering yang lebih baik jika dibandingkan dengan PCA (Principal Component Analysis) yaitu dengan tingkat SSE (Sum of Squared Error) yang lebih kecil.

Kata Kunci : Data mining, ICA, PCA, data numerik, dimensi tinggi, independent,

---

### Abstract

Data mining is defined as a techniques of solving data warehouse problem, in order that data collection which doesn't have anything information can be mined and analyzed. Data collection can't be processed directly, data with high dimensionality must be preprocessed so the output is truly maximal when data is mined.

ICA (Independent Component Analysis) which can minimize dimension and then new dimension is named IC (Independent Component) is a reduction of numeric data. Reduction in ICA is finding dimension which is independent each other to be a new dimension. This condition is stronger than finding uncorrelation between dimension.

Curse of dimensionality phenomenon can be solved by minimizing dimension using ICA.

Evaluation using ICA can reduce data colon tumor with 2000 dimension extracted to be 60 IC and DLBCL dataset with 4026 dimension extracted to be 46 IC. Both datasets have the best clustering evaluation in one IC. As seen in the value of SSE in data colon tumor is 0 and the value of SSE in DLBCL is 0.97. ICA gives better clustering evaluation than PCA (Principal Component Analysis) because ICA has smaller SSE values than PCA.

Keywords : Data mining, ICA, PCA, numeric data, high dimension, independent,

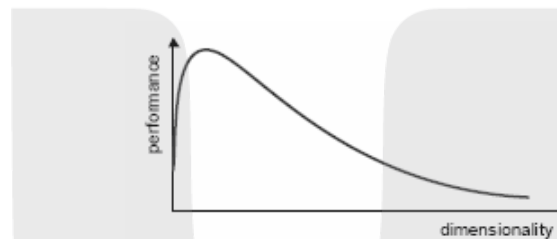
---

# 1. Pendahuluan

## 1.1 Latar Belakang

*Data mining* merupakan suatu bidang yang sangat berkembang pesat karena adanya kebutuhan untuk memanfaatkan gudang data yang dimiliki, yaitu dengan melakukan serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Ada beberapa tantangan dalam *data mining* yaitu salah satunya *dimensionality* "fenomena *curse of dimensionality*".

Fenomena *curse of dimensionality* merupakan fenomena yang terjadi pada data mining dengan adanya peningkatan jumlah dimensi data sehingga dapat menyebabkan turunnya performansi. Oleh karena itu harus dilakukan *preprocessing* data yaitu dengan mengurangi dimensi tersebut [15].



Gambar 1-1: Ilustrasi Curse of Dimensionality

Tujuan utamanya antara lain menghindari *curse of dimensionality*, mengurangi jumlah waktu dan memori yang diperlukan oleh algoritma data mining, membuat data lebih mudah divisualisasikan dan membantu mengeliminasi data yang tidak relevan dengan begitu maka data akan menjadi lebih akurat.

Terdapat beberapa teknik yang sering dipakai dalam mereduksi dimensi diantaranya *principle component analysis* (PCA) dan teknik yang akan penulis analisis yaitu *independent component analysis* (ICA) yang merupakan perluasan dari PCA itu sendiri. ICA melakukan proses pengurangan dimensi dengan melihat keterhubungan antar dimensi dan independensi secara statistik. Oleh karena itu penulis tertarik untuk menganalisa sejauh mana ICA bekerja, membandingkan hasil evaluasi *clustering* dengan ICA dan PCA serta menganalisis apakah ICA dapat memberikan hasil yang lebih baik dalam mereduksi dimensi, juga sejauh mana pengaruhnya terhadap hasil *clustering*.

## 1.2 Perumusan masalah

Berdasarkan latar belakang terdapat permasalahan tersebut, maka akan dilakukan analisa lebih lanjut. Diantaranya bagaimana mendapatkan hasil yang terbaik pada saat melakukan mining dengan data yang memiliki dimensi tinggi. Diharapkan dengan algoritma ICA yaitu dengan menganalisa *statistical independent component* dapat mengurangi dimensi secara optimal dan mempengaruhi hasil

*clustering* dengan akurasi yang baik. Pengukuran tingkat akurasi dan optimal dapat diukur diantaranya dengan SSE, *overall entropy*, dan overall purity.

### 1.3 Tujuan

Berdasarkan rumusan masalah di atas, maka tujuan dari tugas akhir ini adalah:

1. Mengimplementasikan ICA untuk mereduksi dimensi data .
2. Menganalisis ICA terkait dengan hasil reduksi .
3. Menganalisis hasil evaluasi *clustering* setelah dilakukan pengurangan dimensi dengan ICA.
4. Membandingkan hasil evaluasi *clustering* ICA dan PCA.

### 1.4 Batasan Masalah

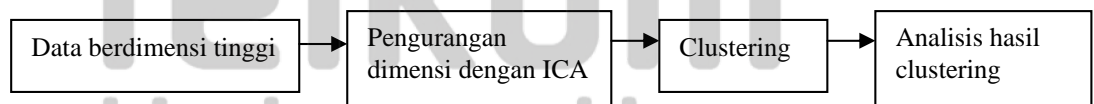
Adapun batasan masalah yang digunakan dalam penelitian ini antara lain :

1. Dataset yang digunakan berupa data numerik .
2. Menganalisis data yang telah direduksi ICA dengan menggunakan *tools* clementine 10.0 .
3. Teknik data mining yang dipakai *K-means clustering*.

### 1.5 Metodologi penyelesaian masalah

Metode yang digunakan dalam penyelesaian tugas akhir ini dengan langkah kerja sebagai berikut :

1. Studi Literatur :  
Pencarian referensi dan mempelajari berbagai literatur yang layak yang berhubungan dengan *data mining* , *clustering*, *preprocessing*, *dimensional reduction*, clementine 10.0, PCA dan ICA .
2. Melakukan analisis dan perancangan perangkat lunak  
Sebelum melakukan analisis terdapat beberapa tahapan secara umum, diantaranya pengurangan dimensi dengan ICA untuk data berdimensi tinggi kemudian dilakukan *clustering* dan analisis hasil *clustering* pada data tersebut. Hasil pengurangan dimensi dapat diuji setelah dilakukan *clustering* sehingga dapat dianalisis pada dimensi-dimensi tertentu akan menghasilkan klasterisasi yang optimal .



Gambar 1-2: Tahapan Umum

3. Melakukan implementasi algoritma ICA dengan matlab versi 7 .
4. Melakukan pengujian dan analisis terhadap hasil implementasi ICA dengan menggunakan clementine 10.0 .
5. Pengambilan kesimpulan dan penyusunan laporan tugas akhir.

## 5 Kesimpulan dan Saran

### 5.1 Kesimpulan

Dari analisis hasil pengujian pada bab 4 dapat disimpulkan :

1. Fenomena *Curse of Dimensionality* dapat diselesaikan dengan pengurangan dimensi dengan ICA .
2. ICA sangat baik digunakan untuk melakukan preprocessing data mining yaitu pada pengurangan dimensi .
3. Pada dataset *colon tumor* yang memiliki 2000 dimensi dapat diringkas menjadi 61 IC. Berdasarkan hasil analisis *clustering*, evaluasi sebelum reduksi sebesar 463.58 sedangkan evaluasi setelah reduksi pada 61 IC yaitu dengan nilai SSE sebesar 41.175. Pada data set DLBCL, data asli yang memiliki 4026 dimensi dapat diringkas menjadi 46 IC dan berdasarkan hasil *analisis clustering*, evaluasi sebelum reduksi sebesar 625.33 sedangkan evaluasi terbaik setelah reduksi pada 46 IC yaitu dengan nilai SSE sebesar 22.220.
4. Dari analisis hasil *K-Means Clustering* pada data set *colon tumor* dan DLBCL diperoleh hasil bahwa ICA dapat memberikan pengaruh performansi yang lebih baik dari pada PCA untuk *K-Means Clustering* pada kedua data set .

### 5.2 Saran

1. Membandingkan ICA dengan algoritma factor analysis
2. Mengimplementasikan ICA untuk algoritma *clustering* selain *K-Means Clustering*
3. Diperlukan penelitian sebelumnya untuk menentukan nilai eigenvalues pada data *colon tumor* dan DLBCL

Telkom  
University

## Daftar Pustaka

- [1] Cheung Samson."Dimension Reduction".2004
- [2] Data set *colon tumor* dan DLBCL. Kent Ridge Bio-medical Data Set Repository. Tersedia: <http://sdmc.lit.org.sg/GEDatasets/Datasets.html>
- [3] Delorme,Arnaud ,"ICA (Independent Component Analysis) for dummies".  
<http://www.sccn.ucsd.edu/~arno/indexica.html> . [11 Desember 2007]
- [4] Fodor ,Imola K."A survey of dimension reduction techniques"2002
- [5] Guido Sanguinetti."Dimensional reduction of clustered data sets"2007
- [6] Independent component analysis.  
[http://en.wikipedia.org/wiki/Independent\\_component\\_analysis](http://en.wikipedia.org/wiki/Independent_component_analysis),  
[11 Desember 2007]
- [7] Jaymie Strecker."Independent Component Analysis (ICA)"2006
- [8] Jiawei Han and Micheline Kamber. "*Data Mining : Concepts and Techniques*", Intelligence Database Systems Research Lab, School of Computing Science, SIMO Fraser University.
- [9] Leskovec,Jure.(2006).Dimensionality reduction PCA, SVD, MDS, ICA, and friends.[http://www.cs.cmu.edu/~guyton/Class/10701-S06/Handouts/recitations/recitation-pca\\_svd.ppt](http://www.cs.cmu.edu/~guyton/Class/10701-S06/Handouts/recitations/recitation-pca_svd.ppt) [11 Desember 2007]
- [10] Lieven De Lathauwer, Bart De Moor and Joos Vandewalle ,"An introduction to independent component analysis"
- [11] M Lennon,G Mersier,M.C.Moucho,L.Hubert-Moy,"Independent component analysis as a tool for the dimensionality reduction and the representation of hyperspectral images"2001
- [12] Nhita ,Fhira.(2007).Analisis principal component analysis pada unsupervised learning untuk data berdimensi tinggi (Tugas Akhir).
- [13] Oursland Alan, Mahmood Nasim, De Paula Judah." Case Studies of Independent Component Analysis"
- [14] Penyiapan Data Preprocessing. Tersedia :<http://lecturer.eepis-its.edu/~tessy/lecturenotes/datamining/chapter2.pdf>.[20 Agustus 2008]
- [15] Pramudiono Iko."Menambang Permata Permai di Gunung Data ".2003

- [16] Ravisekar Bharat." A Comparative Analysis of Dimensionality Reduction Techniques".2006
- [17] Ruusunen Mika and van Ast Jelmer." A Guide to Independent Component Analysis -Theory and Practice".2004
- [18] Smith, Lindsay I. (2002). A tutorial on Principal Components Analysis. Tersedia:[http://csnet.otago.ac.nz/cosc453/student\\_tutorials/principal\\_components.pdf](http://csnet.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf). [19 mei 2008]
- [19] Sudaryatno, Bambang. Analisis Faktor-faktor yang Mempengaruhi Mahasiswa dalam Memilih Perguruan Tinggi Sekolah Tinggi Manajemen Informatika dan Komputer "AMIKOM"Yogyakarta.Tersedia:<http://dosen.amikom.ac.id/downloads/artikel/Jurnal%20upload.doc>. [18 Agustus 2008]
- [20] Suharto, Toto. Materi Perkuliahan Rekayasa Perangkat Lunak (RPL). STT Telkom: Bandung
- [21] Teori Analisis Korelasi Mengenal Analisis Korelasi. Tersedia <http://www.jonathansarwono.info/korelasi/korelasi.htm>. [7 agustus 2008]
- [22] Variansi. Tersedia:<http://id.wikipedia.org/wiki/Varians>. [28 juli 2008]
- [23] Wei,Jinpeng."Unsupervised Learning and Dimensionality Reduction"