

# PENGUNAAN ALGORITMA KLASIFIKASI DALAM DATA MINING

A'ang Subiyakto

Program Studi Sistem Informasi Fakultas Sains dan Teknologi UIN Jakarta

Jl. Ir. H. Juanda No. 95 Ciputat Tangerang Banten 15412

E-mail: aangsubiyakto@yahoo.com

## Abstrak

*Paper ini mengkaji hasil penelitian tentang kecenderungan penggunaan metode algoritma dalam pengembangan data mining (DM). Selain itu juga menjelaskan model framework konseptual dari DM dalam workflow untuk lebih memberikan gambaran DM lebih jelas. Kedua penjelasan tentang DM tersebut bertujuan untuk menunjukkan peranan algoritma dalam framework dan workflow DM sebagai tools. Kesimpulannya adalah bahwa metode algoritma pengklasifikasian secara teoritik menjelaskan beberapa kelebihan dalam pengembangan DM. Hal ini didukung oleh hasil penelitian bahwa metode algoritma pengklasifikasian data masih menjadi pilihan terkait simplicity, elegance dan robustness pengembangan DM..*

**Kata Kunci:** DM, algoritma pengklasifikasian, simplicity, elegance dan robustness

## 1. Pendahuluan

Salah satu konferensi internasional tentang DM yaitu *International Conference on Data Mining (ICDM)* yang diselenggarakan oleh *Institute of Electrical and Electronic Engineers (IEEE)* pada 21 Desember 2006 mempresentasikan sebuah makalah hasil penelitian survey oleh sebuah tim yang beranggotakan 8 (delapan) orang peneliti dari universitas-universitas di Amerika, Inggris, Australia dan China. Penelitian ini mengidentifikasi penggunaan algoritma dalam DM. Survey dilakukan terhadap 18 nominasi di *Association for Computing Machinery (ACM) Knowledge Discovery in Databases (KDD) Innovation Award* and *IEEE ICDM Research Contributions Award* dalam 10 (sepuluh) area topik, meliputi: 1) *association analysis*, 2) *classification*, 3) *clustering*, 4) *statistical learning*, 5) *bagging and boosting*, 6) *sequential patterns*, 7) *integrated mining*, 8) *rough sets*, 9) *linkmining* dan 10) *graph mining* [1].

Berikut ini hasil dari penelitian tersebut bahwa pada 10 (sepuluh) urutan teratas algoritma yang sering digunakan dalam DM, yaitu : 1) C45, 2) K-Means, 3) SVM, 4) Apriori, 5) EM, 6) PageRank, 7) AdaBost, 8) kNN, 9) Naive Bayes dan 10) CART. AdaBost, kNN, dan Naive Bayes memperoleh jumlah voting yang sama. Ada yang menarik

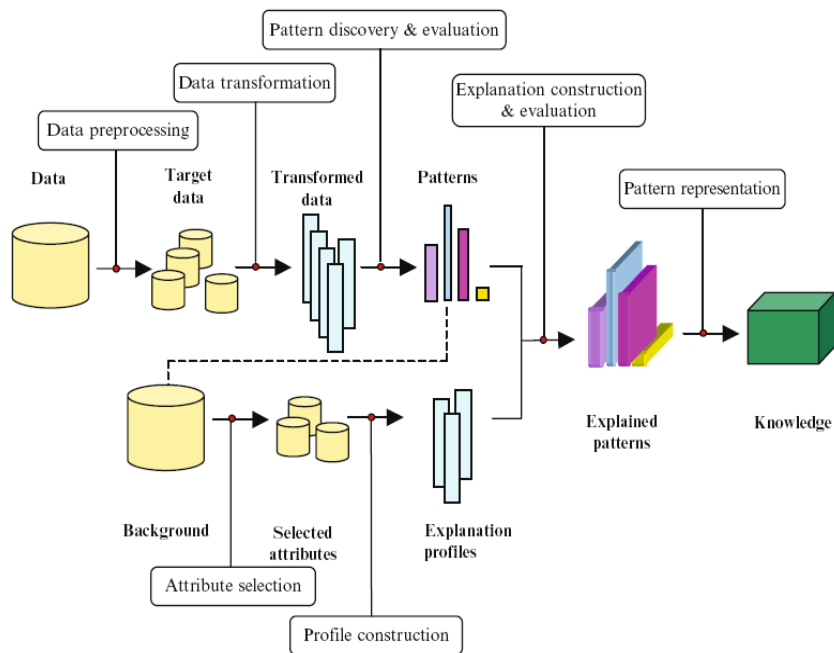
dari hasil ini. Semua nominasi 4 (empat) algoritma yaitu algoritma C45, *k-Nearest Neighbor Classification* (kNN), Naive Bayes, *Classification and Regression Trees* (CART) dari area klasifikasi data masuk dalam sepuluh nominasi. Paper ini menjelaskan tentang gambaran hasil penelitian survey dari penggunaan metode algoritma dalam DM khususnya metode algoritma pengklasifikasian data.

## 2. Konsep Proses Permodelan DM

Hornick MF *et al.* [2] mendefinisikan DM sebagai sebuah proses menemukan model dan relasi-relasi dalam data. Sebuah model yang menggambarkan penggunaan data secara historical dan mengaplikasikannya dalam suatu model baru untuk memperkirakan kecenderungan tertentu (*classification* dan *regression*), segmentasi populasi (*clustering*), penentuan relasi dalam populasi (*association*) dan sebagai identifikasi identitas (*attribute importance*).

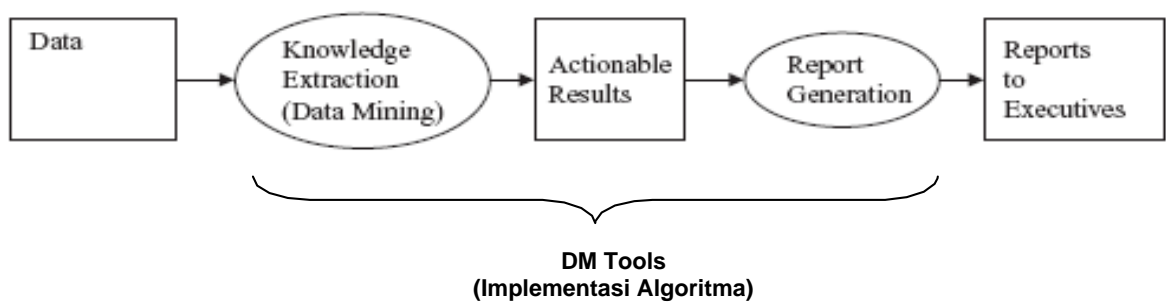
Yao *et al.* [3] menjelaskan sebuah *framework* konseptual dari DM dalam dua isu, yaitu bagaimana menjawab pertanyaan-pertanyaan ilmiah ke dalam bentuk konsep penerapan dan ruang lingkup metode- metodenya. Selanjutnya mereka menjelaskan *framework* konseptual DM dalam tiga *layer* (Gbr. 1), yaitu 1) *philosophy layer*, bagian yang menginvestigasi dasar-dasar pengetahuan DM dan menerjemahkannya ke dalam kenyataan, 2) *technique layer*, menjabarkan pengetahuan ke konteks ilmu komputer dalam bentuk bahasa pemrograman dan 3) *application layer*, menjabarkan efektifitas penggunaan pengetahuan. *Layer* ini fokus pada penerjemahan kosa kata *usefulness* dan *meaningfulness* ke dalam bidang aplikasi tertentu dengan atribut-atribut *efficiency*, *optimization*, *reliability*, *cost-effectiveness* dan *appropriateness* untuk memenuhi kebutuhan riil dalam penerapannya.

Penjelasan *workflow* DM oleh Hornick MF *et al.* [2] memperjelas gambaran *framework* di atas (Gbr. 2). Secara sederhana mereka menjelaskan DM sebagai sebuah aliran proses sebagai berikut: *pertama*, mendefinisikan masalah dan sasaran DM, mengidentifikasi kebutuhan data dan menentukan kualitas data yang dibutuhkan.



**Gbr. 1.** Tiga Layer Framework Konseptual DM [3]

*Kedua*, mentransformasikan data dengan DM tool dengan pemanfaatan algoritma ke dalam bentuk model data baru. Peranan algoritma dalam sub proses ini adalah mengekstraksi data sumber menjadi model data baru sesuai kebutuhan domain dan *ketiga*, pengolahan data dari model data baru ke dalam bentuk informasi sesuai jenis kebutuhan proses bisnis pengguna.



**Gbr. 2.** Workflow DM ([2], Diadopsi)

### 3. Metode Algoritma Klasifikasi

Penelitian ini menyebutkan 4 (empat) metode algoritma klasifikasi yang cenderung digunakan dalam pengembangan DM, yaitu: 1) C45, metode ini menjadi pilihan pertama yang sering digunakan dalam pengembangan DM karena kecepatan dalam pengklasifikasian pohon keputusan disamping dapat mengkonstruksi pengklasifikasian dengan aturan-aturan yang lain [1]. 2) *k*-NN, beberapa hal yang menjadi perhatian dalam penggunaan algoritma ini adalah penggunaan pilihan *k*, jika *k* sangat kecil maka akan mengakibatkan noise. Sebaliknya jika terlalu besar dapat menyebabkan *N* dengan banyak kelas yang harus diklasifikasikan. Tetapi kesederhanaan metode menjadi nilai lebih sehingga menjadi pilihan banyak developer DM selain itu, algoritma ini mudah untuk dipahami dan diimplementasikan dalam tekniknya. Penelitian ini menyatakan bahwa banyak peneliti berpendapat bahwa algoritma ini lebih baik dari SVM berdasarkan skema pengklasifikasiannya [1].

3) Naive Bayes, penelitian tersebut menyimpulkan bahwa metode algoritma Naive Bayes memiliki keunggulan untuk pengembangan DM, yaitu kemudahan konstruksinya dan tidak membutuhkan parameter skema pengulangan yang kompleks sehingga mudah dalam membaca data dalam jumlah yang besar. Hal ini terjadi karena desain rancangan penuntunan klasifikasi terhadap data. Selain itu, metode ini dinyatakan sebagai algoritma yang mempunyai sifat *simplicity*, *elegance* dan *robustness*. 4) CART, penerapan metode algoritma ini banyak digunakan dalam berbagai bidang yang membutuhkan pengolahan data yang komprehensif. Hanya saja mekanismenya terdiri dari beberapa tahap yang bertingkat meliputi *automatic class balancing*, *automatic missing*, *value handling* dan *allows for cost-sensitive learning*, *dynamic feature construction* dan *probability tree estimation* sehingga tingkat kompleksitas menjadi pertimbangan para peneliti pemula. Hasil akhirnya adalah gambaran atribut berdasarkan prioritas kebutuhan proses.

### 4. Kesimpulan

Berdasarkan hasil penelitian Wu Xindong *et al.* Tentang kecenderungan penggunaan berbagai metode algoritma dalam DM dan penjabaran *framework* konseptual DM dari Yao *et al* serta penjelasan konsep-konsep DM dari Hornick MF *et al.* di atas. Hal ini menunjukkan bahwa algoritma merupakan DM *tools* yang banyak digunakan dengan

metode algoritma klasifikasi merupakan *task* DM yang paling umum dan paling sering dilakukan. Kemudian, semua nominasi area *statistical learning* (SVM dan EM) juga masuk dalam 10 ranking teratas. Topik-topik lanjut seperti *sequential patterns*, *integrated mining*, *rough sets*, *graph mining* memperlihatkan masih kurang terlalu populer. Hanya PageRank dari area link mining yang masuk dalam nominsi ini. Hal ini mungkin karena keberhasilan penggunaannya oleh Google dan yang terakhir, algoritma-algoritma teratas ini merupakan algoritma yang banyak dipakai tidak hanya dalam DM saja. Jika dihubungkan dengan kompetisi-kompetisi DM yang pernah dilakukan, umumnya para pemenangnya menggunakan algoritma-algoritma ada pada daftar ini. Algoritma-algoritma seperti SVM, dan Naive Bayes sangat seringkali digunakan.

#### **Referensi:**

- [1] Wu Xindong *et al.* *Top 10 Algorithms in Data Mining*. Di dalam: *Knowledge Information System*. Vol. 14. London: Springer; 2008. hlm. 1–37
- [2] Hornick MF *et al.* *Java Data Mining: Strategy, Standard, and Practice. A Practical Guide for Architecture, Design, and Implementation*. San Francisco: Morgan Kaufmann Publishers; 2007
- [3] Yao Yiyu *et al.* *A Conceptual Framework of Data Mining*. Di dalam: Lin Y, Xie Y, Wasilewska A, Liau CJ, editor. *Data Mining: Foundations and Practice*. Vol. 118. Berlin: Springer; 2008. hlm. 501-516