

## CONTRAST SET MINING MENGGUNAKAN SUBGROUP DISCOVERY ( CONTRAST SET MINING THROUGH SUBGROUP DISCOVERY )

Wiwin<sup>1</sup>, Shaufiah<sup>2</sup>, Arie Ardiyanti Suryani<sup>3</sup>

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

---

### Abstrak

Teknologi data mining merupakan suatu teknologi untuk menambang atau mengekstraksi pengetahuan menggunakan kumpulan data yang sangat besar. Salah satu pendekatan alternatif dalam data mining untuk mencari perbandingan dari beberapa grup perbandingan disebut contrast set mining. Contrast set mining dapat dilakukan menggunakan subgroup discovery. Salah satu algoritma dari subgroup discovery adalah APRIORI-SD. Berdasarkan hasil pengujian, diperoleh rule-rule yang dapat digunakan untuk mengklasifikasikan data dan diperoleh juga tingkat akurasi dari rule tersebut.

Kata Kunci : data mining, contrast set mining, subgroup discovery,apriori-sd,rule

---

### Abstract

Data mining technology can be used for mining and extracting knowledge with huge amount of data. An alternative approach in data mining for comparing of multiple groups comparison is called contrast set mining. Contrast set mining can be done by using subgroup discovery. One of the subgroup discovery algorithm is the Apriori-SD. Based on implementation of this algorithm, rules can be gained and tested for classifying new data and also rule accuracy can be calculated based on this testing.

Keywords : data mining, contrast set mining, subgroup discovery,apriori-sd,rule

---

Telkom  
University

# BAB I

## PENDAHULUAN

### 1.1 Latar belakang

Perkembangan teknologi dalam bidang basis data pada masa sekarang ini semakin meningkat. Kemajuan teknologi tersebut dapat memungkinkan suatu perusahaan atau organisasi untuk mengumpulkan data dari berbagai sumber dengan mudah dan cepat. Kemudahan dalam mengumpulkan data ini membuat volume data semakin besar dan terus bertambah. Semakin besar volume data yang ada, semakin besar informasi dan pengetahuan yang dapat didapat dari data tersebut. Data tersebut dapat dianalisis dan diinterpretasi secara manual, namun analisis data secara manual tersebut bersifat lambat dan mahal.

Kebutuhan analisis data tersebut dapat dijawab dengan adanya teknologi data mining, yaitu suatu teknologi untuk menambang atau mengekstraksi pengetahuan menggunakan kumpulan data yang sangat besar. Data mining juga mampu menjadi solusi untuk mendapatkan pola informasi pada data yang berjumlah besar. Salah satu pendekatan alternatif dalam data mining untuk mencari perbandingan dari beberapa grup perbandingan disebut *contrast set mining*. *Contrast set mining* memanfaatkan *contrast set* dalam mendeteksi antar kelompok data yang berbeda. *Contrast set* merupakan keterhubungan antara pasangan atribut dan nilai pada suatu data set. *Contrast set mining* dapat dilakukan menggunakan *decision tree*, *rule learning*, dan *subgroup discovery*.

*Subgroup discovery* didefinisikan sebagai berikut diberikan sebuah populasi dari individu-individu dan properti dari individu tersebut yang paling dianggap menarik, kemudian temukan populasi subgroup yang secara statistik "paling menarik". Sebuah populasi subgroup dikatakan "paling menarik" apabila mudah dimengerti manusia, valid pada data testing atau data baru dengan suatu derajat kepastian, dan biasanya memvalidasi beberapa hipotesis yang dicari kebenarannya.

Subgroup discovery cocok diterapkan dalam menganalisis data karena subgroup discovery dapat menghasilkan deskripsi singkat dan mudah dimengerti dari hasil analisis data yang dilakukan. Standar klasifikasi dari algoritma *rule learner* tidak dapat langsung digunakan dengan baik dalam subgroup discovery. Oleh karena itu, digunakanlah algoritma yang dapat diadaptasikan untuk mengimplementasikan *subgroup discovery*, salah satunya adalah APRIORI-SD yang merupakan adaptasi dari algoritma APRIORI yang digunakan untuk membangun rule klasifikasi dalam *association learning*. Beberapa adaptasi dari algoritma APRIORI diperlukan untuk menjalankan tugas klasifikasi, salah satunya adalah dengan menerapkan suatu skema pembobotan pada saat proses pemilihan rule yang dihasilkan.

Berdasarkan hal tersebut, maka pada Tugas Akhir ini akan dilakukan contrast set mining menggunakan subgroup discovery dengan menggunakan algoritma APRIORI-SD.

## 1.2 Perumusan masalah

Dengan mengacu pada latar belakang diatas, maka rumusan permasalahan yang akan dibahas dan diteliti pada Tugas Akhir ini adalah:

1. Menentukan rule set dari data yang telah ditentukan
2. Menerapkan pengujian dan analisis terhadap hasil yang diperoleh dari *subgroup discovery*

Adapun batasan masalah yang digunakan adalah :

1. Data yang akan diolah dalam pengerjaan Tugas Akhir ini adalah data jamur dan data *nursery* yang berasal dari *UCI Datasets Repository* serta data SMBB calon Mahasiswa IT Telkom tahun 2005,2006, dan 2007
2. Algoritma yang diterapkan dalam menentukan *rule set* adalah algoritma APRIORI-SD
3. Data set yang akan digunakan telah terbebas dari noise

## 1.3 Tujuan

Tujuan dari pengerjaan Tugas Akhir ini adalah :

1. Menerapkan algoritma APRIORI-SD untuk menghasilkan *rule set* pada *subgroup discovery*
2. Mengimplementasikan dan menganalisis nilai support serta hasil rule set yang dihasilkan dan menerapkan pengujian pada data set yang baru

## 1.4 Metodologi penyelesaian masalah

Metodologi yang diterapkan dalam menyelesaikan masalah pada Tugas Akhir ini yaitu :

1. Studi Literatur  
Mempelajari dan memahami konsep mengenai *Subgroup Discovery*, algoritma dan metode pendukung yang akan diterapkan di Tugas Akhir ini melalui berbagai literatur seperti jurnal, buku dan makalah dari berbagai media, khususnya internet
2. Pengumpulan Data  
Mengumpulkan dan mempelajari dokumen / data yang akan dipergunakan dalam pengerjaan Tugas Akhir. Dalam hal ini data yang dimaksud adalah data jamur yang diperoleh dari *UCI Dataset Repository*
3. Perancangan  
Melakukan analisis dan perancangan pengujian yang akan dilakukan dalam pengerjaan Tugas Akhir ini, seperti menentukan diagram *workflow* sistem, perancangan UML diagram, seperti *use case* diagram dan skenarionya, *sequence* diagram, *class* diagram, dan *activity* diagram.
4. Implementasi  
Mengimplementasikan hasil rancangan yang telah dibuat dengan menggunakan bahasa pemrograman Java.
5. Analisis Hasil  
Melakukan analisis terhadap rule set yang dihasilkan, kemudian menguji hasil rule set tersebut dengan data input yang baru.

6. Pengambilan Kesimpulan dan Penyusunan Laporan Tugas Akhir.  
Pembuatan Laporan Tugas Akhir yang mendokumentasikan tahap-tahap kegiatan dan hasil *subgroup discovery* yang dihasilkan dalam Tugas Akhir ini.



## BAB V PENUTUP

### 5.1 Kesimpulan

Dari hasil pengujian dan analisis pada bab sebelumnya, maka kesimpulan yang dapat diperoleh dari Tugas Akhir ini diantaranya :

1. Pada pengujian pada Tugas Akhir ini, berdasarkan pengujian kedua diperoleh data dengan nilai support 20%, confidence 80%, dan  $k = 5$  memiliki tingkat akurasi yang tinggi untuk kedua data, yakni 98,70% untuk data pertama dan 78,9%.
2. Nilai  $k$  tidak berpengaruh terhadap tingkat akurasi dari rule. Tetapi hanya berpengaruh terhadap jumlah rule yang dihasilkan. Akurasi dari rule dipengaruhi oleh banyak sedikitnya data uji yang terklasifikasi dengan benar oleh rule yang dihasilkan.
3. Nilai support dan confidence sangat relatif terhadap data yang digunakan. Parameter support dan confidence berpengaruh terhadap banyak sedikitnya rule yang dihasilkan, lamanya waktu eksekusi aplikasi, serta banyak sedikitnya alokasi memori yang digunakan.

### 5.2 Saran

Dari hasil yang diperoleh dari Tugas Akhir ini, perlu diperhatikan beberapa saran untuk perbaikan dan pengembangan penelitian selanjutnya sebagai berikut:

1. Penelitian selanjutnya sebaiknya dapat menggunakan algoritma subgroup discovery lainnya agar dapat dibandingkan dengan algoritma yang digunakan dalam Tugas Akhir ini.
2. Sebaiknya menggunakan nilai support yang lebih kecil lagi, agar keakurasian dari rule tersebut besar.

## DAFTAR PUSTAKA

- [1] Atzmueller, Martin. *Subgroup Discovery*. 2005
- [2] Bay, Stephen D., Pazzani, Michael J. *Detecting Group Differences: Mining Contrast Sets*. *Data Min. Knowl. Discov.* 5(3): 213-246 (2001)
- [3] Jovanoski, Viktor , Lavrač, Nada. *Classification Rule Learning with APRIORI-C*.
- [4] Kavsek, Branko , Lavrač, Nada. *Apriori-SD : Adapting Assosiation Rule Learning to Subgroup Discovery*. *Applied Artificial Intelligence*, 20:543-583.
- [5] Kralj, Petra, Lavrač, Nada, Gamberger, Dragan, Krstačić, Antonija. *Contrast Set Mining for Distinguishing Between Similar Diseases*. *AIME-2007*.
- [6] Kralj, Petra, Lavrač, Nada, Gamberger, Dragan, Krstačić, Antonija. *Contrast Set Mining Through Subgroup Discovery Applied to Brain Ischaemia Data*.
- [7] Kralj, Petra, Lavrač, Nada, Webb, Geoffrey I. 2009. *Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern, and Subgroup Mining*. *Journal of Machine Learning Research* 10 : 377-403.
- [8] Lavrač, Nada, Gamberger, Dragan, Cestnik, Bojan, Flach, Peter. *Decision Support Through Subgroup Discovery: Three Case Studies and the Lessons Learned*. 2004. Netherland : Kluwer Academic Publishers. *Machine Learning*, 57:115-143.
- [9] Pang Ning Tan, Michael Steinbach and Vipin Kumar. *Introduction to Data mining*
- [10] Utami, Trisnie C. 2009. *Deteksi Karakteristik Antar Kelompok Data Berbeda dengan Menggunakan Contrast Set Mining*. Bandung: IT Telkom.
- [11] [http://en.wikipedia.org/wiki/Confusion\\_matrix](http://en.wikipedia.org/wiki/Confusion_matrix) diakses tanggal 2 Agustus 2010
- [12] [http://www2.cs.uregina.ca/~hamilton/courses/831/notes/confusion\\_matrix/confusion\\_matrix](http://www2.cs.uregina.ca/~hamilton/courses/831/notes/confusion_matrix/confusion_matrix) diakses tanggal 2 Agustus 2010
- [13] <http://www.kaskus.us/showthread.php?t=4057452&page=11> diakses tanggal 30 September 2010