PENGGUNAAN ROUGH SET APPROACH SEBAGAI KRITERIA VARIABLE SELECTION DALAM TASK CLASSIFICATION PADA DATA MINING

Ardedi Frianto Ambarita[1], Moch. Arif Bijaksana[2], Kiki Maulana[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Data mining adalah proses pencarian pola dari kumpulan data. Dalam data mining ada 3 task utama yaitu klasifikasi, asosiasi dan klasterisasi. Klasifikasi merupakan task data mining untuk menemukan pola dari kumpulan data. Tujuan dari pecarian pola tersebut adalah untuk menjawab nilai dari suatu data yang belum diketahui nilainya. Tahapan klasifikasi pada data mining dimulai dengan pembersihan data (preprocessing), pembentukan model dan evaluasi (testing) model. Tahap preprocessing merupakan salah satu tahap yang sangat penting dalam klasifikasi data mining, karena tahap ini merupakan tahap penyediaan data yang digunakan untuk pembentukan pola nantinya. Salah satu bentuk tahap preprocessing dalam data mining adalah pemilihan variabel atau yang sering juga disebut dengan variable selection, variable subset selection, feature selection, feature reduction, atau attribute selection.

Tujuan dari pemilihan variabel adalah untuk mendapatkan data yang optimal, sehingga hasil akhir model yang didapat lebih optimal, selain itu juga untuk mendapatkan model yang lebih ringkas dan mempercepat proses learning. Dalam tugas akhir ini diimplementasikan pemilihan variabel menggunakan Rough Set. Rough Set digunakan untuk mendapatkan perkiraan rule yang singkat dari suatu data, dalam hal ini pengurangan variabel atau kolom.

Kata Kunci : Data mining, Rough Set, Klasifikasi, Preprocessing,Variable selection, Feature selection

Abstract

Data mining is a process of finding a data group pattern. There are 3 tasks in data mining, which are classification, association, and clusterization. Classification is a task of finding the data group pattern. The aim of finding the data group pattern is to get the data value which has not known. The classification task of data mining is started with data pre-processing, model-building, and model-testing.

The pre-processing task is a very important step in classification task of data mining. It is because this is the phase where the data to used in the patternbuilding step is prepared. A kind of pre-processing step in data mining is the variable selection or which are commonly called as variable subset selection, feature selection, feature reduction, or attribute selection.

The aim of variable selection is to find the optimal data, so that the final model is more optimal, and also to find the model which is more briefed and to accelerate the learning process. In this thesis, the variable selection implemented by the rough set. Rough set is used to find the brief rule approximation of a data, in this case the variable or column subtraction.

Keywords : Data mining, Rough Set, Classification, Pre-processing,Variable selection, Feature selection

1. Pendahuluan

1.1 Latar belakang

Perkembangan teknologi dalam hal penyimpanan dan pengolahan data pada saat ini sangat berkembang pesat, hal ini dikarenakan oleh kebutuhan manusia akan informasi yang cepat dan tepat. Oleh itu, pada saat ini sebagian besar organisasi atau perusahaan menyimpan informasi transaksi yang berlangsung di dalam organisasi tersebut, hal ini ini juga didukung oleh murahnya penerapan teknologi. Data yang banyak yang tersimpan di dalam *storage* perusahaan tidaklah berarti apa-apa jika tidak diolah dan digunakan. Kebutuhan akan analsisis data tersebut mendorong penerapan dari berbagai teknik analisis data dari berbagai bidang ilmu seperti statistika, kecerdasan buatan, database dan lain-lain untuk menganalisis data tersebut agar dapat digunakan. Penggabungan analisis dari berbagai bidang ilmu tersebut dikenal dengan *data mining*.

Data mining bertujuan untuk menemukan pola yang ada dalam suatu kumpulan data. Pola yang didapat dalam *data mining* dapat digunakan sebagai pertimbangan dalam pengambilan keputusan dalam instansi atau perusahaan sumber data tersebut. Salah satu teknik dalam *data mining* adalah klasifikasi. Klasifikasi merupakan proses menemukan model atau fungsi untuk dapat memprediksi suatu objek yang belum diketahui kelasnya.

Dalam klasifikasi, data yang digunakan untuk proses pembentukan pola sangatlah penting. Untuk mendapatkan pola yang lebih baik dibutuhkan data yang baik pula. Pada klasifikasi ada dua proses utama yaitu proses *training* dan proses *testing*. Proses klasifikasi digunakan varibel input untuk *training* dan *testing* untuk mengetahui tingkat akurasi kebenaran model yang dibangun. Pemilihan variabel yang digunakan untuk proses klasifikasi sangatlah penting. Tujuan pemilihan variabel dalah untuk meningkatkan performansi prediksi suatu kelas, harga prediksi yang efektif, kemudahan visualisasi, pemahanan data dan juga untuk mengurangi dimensionalitas dari variabel input.

Pada tugas akhir ini diimplementasikan pemilihan variabel sebelum proses klasifikasi dengan menggunakan *Rough Set Approach*. *Rough Set Approach* merupakan teknik pendefenisian kelas yang ekivalen dalam *dataset* secara kasar (*roughly*). *Rough Set Approach* bertujuan untuk menemukan minimal subset dari variabel yang akan digunakan pada proses *learning*. Himpunan minimal variabel dari suatu *dataset* dibuat dengan memilih minimal subset dari *dataset* asal. Himpunan variabel yang dapat mewakili beberapa *rule* dinamakan *core*. Setelah seluruh *core* yang dapat mewakili *rule* yang lain didapatkan, maka *dataset* yang baru didapatkan dengan menghapus variabel yang tidak termasuk dalam *core*, setelah akan didapat *dataset* baru yang sebagai hasil dari *variable selection.*

1.2 Perumusan masalah

Klasifikasi dalam *data mining* digunakan untuk mencari pola dengan cara menganalisis sekumpulan *dataset* yang mendeskripsikan dan membedakan kelas-kelas data. Tujuan dari pembentukan pola adalah untuk memprediksi data yang

belum diketahui kelasnya. Pada klasifikasi, *dataset* akan dijadikan sebagai data input untuk proses *training* dan *testing*. Namun untuk jumlah data yang besar dan variabel yang banyak, sangat kompleks untuk membuat model *classifier* dari *dataset* tersebut. Oleh karena itu, sebelum proses *training* pada klasifikasi, variabel input yang digunakan perlu dipilih untuk mendapatkan pola yang lebih baik, sehingga dimensionalitas data dapat dikurangi.

Pemilihan variabel yang dimaksudkan adalah memilih beberapa variabel yang ada, sehingga tidak semua variabel dalam *dataset* digunakan. Pada tugas akhir ini dianalisis penggunaan *Rough Set Approach* sebagai kriteria pemilihan variabel yang akan digunakan dalam klasifikasi pada *data mining*.

Dari uraian diatas, dirumuskan masalah dalam tugas akhir ini adalah :

1. Bagaimana jika jumlah variabel *dataset* terlalu banyak (*resource* yang dibutuhkan sangat besar pada saat *learning*).
2. Pemilihan variabel digunakan sebelum proses *learning* untuk mengurangi dimensionalitas *dataset.*

Rough Set Approach merupakan teknik untuk mengurangi atau menghapus suatu variabel yang mungkin *redundant*, tanpa menghilangkan informasi penting untuk pembuatan *classifier*. Hal ini berdasarkan konsep himpunan *upper* dan *lower approximation, approximation space* dan himpunan model.

Dalam Tugas Akhir ini, yang dibahas adalah implementasi penggunaan *Rough Set Approach* sebagai kriteria *variable selection* dalam *task classification* pada *Data Mining*, dengan batasan masalahnya sebagai berikut:

1. *Dataset* yang digunakan adalah *dataset* yang bertipe data record.
2. *Dataset* yang digunakan sudah dalam bentuk data diskret dan tidak menangani data kontinu.
3. Tidak menangani *missing value* dan *data cleaning*.
4. *Dataset* yang akan digunakan sebagai data *training* dan data *test* sudah tersedia.
5. Implementasi hanya pada perangkat lunak pemilihan variabel dari *dataset*, sedangkan untuk pembentukan *classifier* menggunakan perangkat lunak yang telah ada yaitu Weka.

1.3 Tujuan

Tujuan dari pembuatan tugas akhir ini adalah sebagai berikut :

1. Merancang dan membangun perangkat lunak untuk pemilihan variabel yang akan digunakan pada *task classification* dengan menggunakan *Rough Set Appoarch*.
2. Membandingkan hasil pemilihan variabel dengan perangkat lunak yang sudah ada, yaitu Weka dan Clementine.
3. Membandingkan hasil pengukuran *classifier* sebelum dan sesudah dilakukan *variable selection* menggunakan *Rough Set Appoarch* pada *dataset* yang meliputi *precison*, *recall* dan *accuracy* yang dihitung dari *confusion matrix* yang dihasilkan oleh Weka.

1.4 Metodologi penyelesaian masalah

Metodologi yang digunakan untuk menyelesaikan tugas akhir ini adalah:

1. Studi Literatur
   Pada tahap ini akan dilakukan pendalaman materi, identifikasi masalah dan metodologi yang akan digunakan dalam pemecahan masalah. Mempelajari teori dasar proses pemilihan variabel secara umum, teori dasar *Rough Set Approach* serta penggunaanya pada pemilihan variabel, cara pemilihan *core* yang dapat mewakili *rule* dari suatu *dataset,* pengukuran *classifier* berdasarkan *confusion matrix* yang dihasilkan oleh *software* klasifikasi yang digunakan. Setelah melakukan studi literatur akan dilakukan pengujian apakah metode *Rough Set Approach* dapat digunakan sebagai proses pemilihan variabel pada *data mining*.
2. Perancangan
   Pada tahap ini dilakukan perancangan sistem yang akan dibangun, yang meliputi analisis terhadap kebutuhan perangkat lunak baik itu input maupun output sistem yang dibangun. Setelah itu dibuat diagram aliran data yang berfungsi untuk memberikan pennggambaran aliran data dan cara kerja sistem. Kemudian dilakukan pengumpulan data.
3. Implementasi
   Impementasi pada program untuk pemilihan variabel dari *dataset* dengan menggunakan *Rough Set Approach* dimana variabel hasil pemilihan variabel akan digunakan pada klasifikasi di Weka untuk membentuk model.
4. Pengujian
   Pada tahap ini dilakukan pengujian terhadap sistem yang dibangun, serta melakukan perbaikan terhadap *bug* dan *error* yang ditemukan pada perangkat lunak yang dibangun. Setelah dilakukan pengujian terhadap sistem, maka dilakukan proses pemilihan variabel pada *dataset* yang ada menggunakan sistem yang dibangun dan hasilnya akan digunakan pada klasifikasi di Weka.
5. Analisis dan pembuatan laporan
   Melakukan analisis terhadap hasil pemilihan variabel sistem yang dibangun juga untuk hasil pemilihan variabel menggunakan Weka dan Clementine. Dari hasil pemilihan variabel yang dibangun, Weka dan Clementine akan dianalisis model yang dihasilkan dengan menggunakan weka yang meliputi parameter yang telah ditentukan sebelumnya. Setelah itu dilakukan pembuatan laporan hasil analisis.

5. Kesimpulan dan Saran

5.1 Kesimpulan

1. *Variable slelection* menggunakan *Rough Set* menghasilkan rata-rata akurasi paling tinggi dari 3 metode *variable selection* yang digunakan.
2. *Variable slelection* menggunakan *Rough Set* menghasilkan rata-rata akurasi yang lebih tinggi dari rata-rata akurasi *dataset* awal.
3. Model yang dihasilkan dari *dataset* hasil *variable slelection* menggunakan *Rough Set* dapat menyamai, bahkan lebih ringkas dari model yang dihasilkan dari *dataset* awal..
4. *Variable slelection* menggunakan *Rough Set* menghasilkan rata-rata *precison* paling tinggi dari 3 metode *variable selection* yang digunakan.
5. *Variable slelection* menggunakan *Rough Set* menghasilkan rata-rata *precison* yang lebih tinggi dari rata-rata *precison dataset* awal.
6. *Variable slelection* menggunakan *Rough Set* dapat mengimbangi rata-rata *precison* dari 3 metode *variable selection* yang digunakan, yaitu Clementine 0.941, *Rough Set* 0.933 dan Weka 0.892.
7. *Variable slelection* menggunakan *Rough Set* dapat mengimbangi rata-rata *precison* dari *dataset* awal, yaitu *dataset* awal 0.951 dan *Rough Set* 0.933.
8. Pemilihan variabel menggunakan *Rough Set* dapat dikatakan berhasil, karena dapat mengimbangi rata-rata akurasi, *precision* dan *recall* serta model dari *dataset* awal.

5.2 Saran

1. Perlu diteliti lebih lanjut pendekatan khusus tentang penggunaan *Rough Set* dalam *variable selection*, agar jumlah *record* yang bisa ditangani tidak terbatas.
2. *Rough Set*, selain digunakan pada *filter variable selection* juga dapat *wrapper variable selection* untuk mendapatkan nilai kombinasi variabel yang paling optimal jika kandidat yang didapat lebih dari satu.

Daftar Pustaka

[1] Chouchoulas. A. "A Rough Set Approach to Text Classification", MSc (By Research) in Artificial Intelligence, School of Artificial Intelligence, Division of Informatics, The University of Edinburgh, 1999, http://www.bedroomlan.org/~alexios/files/alexios_msc_thesis.pdf, didownload pada tanggal 26 Januari 2007.

[2] Chouchoulas. A. "Incremental Feature Selection Based on Rough Set Theory", PhD Proposal, Centre for Intelligent Systems and their Applications, Division of Informatics, The University of Edinburgh, 2001, http://www.bedroomlan.org/~alexios/files/alexios_proposal.pdf, didownload pada tanggal 26 Januari 2007.

[3] Clementine 10.0 Node reference

[4] Grzymala-Busse. J.W. " Introduction to Rough Set Theory and Applications", http://www.cit.ac.nz/kes2004/JERZY.pdf, didownload pada tanggal 26 Januari 2007.

[5] Guyon. I. and Elisseeff. A. "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research 3, 2003.

[6] Han. J. and Khamber. M. "Data Mining : Concepts and Techniques", Morgan Kaufmann Publishers, San Francisco,USA, 2001.

[7] http://en.wikipedia.org/wiki/Data_mining

[8] http://en.wikipedia.org/wiki/Feature_selection

[9] http://id.wikipedia.org/wiki/Himpunan

[10] Hui. S. "Rough Set Classication of Gene Expression Data", Department of Computer Science, Faculty of Mathematics, University of Waterloo 2002 .

[11] Kusiak. A. “Rough Set Theory”, Intelligent Systems Laboratory 2139 Seamans Center The University of Iowa, Iowa City, Iowa, http://www.icaen.uiowa.edu/~comp/Public/RoughSets.pdf, didownload pada tanggal 26 Januari 2007.

[12] Pawlak. Z. “Rough Set Elements (1)”, Institute of Theoretical and Applied Informatics Polish Academy of Sciences ul. Baltycka, Gliwice, Poland , http://207.203.212.204/images/articles/RoughSetElements1.pdf, didownload pada tanggal 26 Januari 2007.

[13] Pawlak. Z.,Grzymala-Busse, Slowinski. R., and Ziarko. W. “Rough Sets”, Communication of the ACM November 1995/Vol. 38, No. 11, 1995,http://207.203.212.204/Images/articles/pawlakacm.pdf, didownload pada tanggal 26 Januari 2007.

[14] Pressman Roger. S. “Software Engineering a Practitioner Approach”, McGraw- Hill Inc, Sixth Edition, 2005.

[15] Swiniarski. R.W. and Skowron. A. “Rough set methods in feature selection and recognition“,Pattern Recognition Letters 24 833–849, 2003, http://logic.mimuw.edu.pl/Grant2003/prace/EPRLSkowron1.pdf , didownload pada tanggal 26 Januari 2007.

[16] Tan, Pang-Ning, et all. “Introduction to Data Mining”, Pearson Education, Inc., Boston, 2006.

[17] WEKA Explorer User Guide for Version 3-5-6