ANALISIS DAN IMPLEMENTASI FEATURE SELECTION DENGAN PERPADUAN METODE RAOUGH SETS, MLRELEVANCE CRITERION, DAN PRELEVANCE CRITERION

Kusuma Eka Saputra[1], Toto Suharto[2], Zk. Abdurahman Baizal[3]

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

---

Abstrak

Perkembangan teknologi memacu timbulnya keberagaman data didalamnya. Sedangkan data adalah sumber informasi yang sangat penting. Untuk dapat mengolah data-data tersebut terdapat teknik yang sekarang diimplementasikan, yaitu Knowledge Discovery in Database (KDD). Di dalam KDD terdapat proses data mining yang berutujuan untuk menggali informasi dari data yang ada. Salah satunya dengan cara klasifikasi. Akan tetapi dengan keberagaman data tidak ada jaminan bahwa data itu siap diolah. Contohnya adalah dimenensi data yang begitu besar, hal ini akan menyulitkan dalam proses klasifikasi. Maka dari itu dilakukanlah preprocessing terlebih dahulu.

Preprocessing adalah tahap dimana untuk menyiapkan data agar seefisien mungkin dan terhindar dari noise, missing value, irrelevant feature, redundant feature dll, sehingga diharapkan akan memberikan hasil yang lebih optimal dalam melakukan klasifikasi. Di dalam preprocessing, terdapat salah satu teknik yaitu feature selection. Teknik ini digunakan untuk mengurangi dimensi data atau feature yang dianggap kurang relevan terhadap pemebentukan kelas.

Tugas Akhir ini membahas serta mengimplementasikan teknik feature selection dengan menggunakan metode Rough Sets Theory yang dipadukan dengan MLRelevance Criterion dan PRelevance Criterion. Hasil dari feautre selection dengan menggunakan metode itu, mampu memprediksi feature yang paling relevan. Sehingga tingkat akurasi yang didapatkan mampu mengimbangi precission, recall dan accuracy sebelum dilakukan feature selection.

Kata Kunci : Data mining, preprocessing, klasifikasi, Rough Set, feature selection, variable selection.

Abstract

Advance in technology leads to the emergence of data diversity. Data is an important information source. In order to process the data, there are techniques which can be implemented which is Knowledge Discovery in Database(KDD). In KDD, there are data mining processes to mine information from data. One of the process is classification. Nonetheless, data diversity doesn't guarantee that data is ready to be processed. For example, large data dimention is going to make it difficult for classification task. So, preprocessing must be done.

Preprocessing is a step for preparing data so that the data is efficiently clean from noise, missing value, irrelevant feature, redundant feature, etc thus it will provide optimal result in classification task. In preprocessing, one of the most common method is feature selection.

This thesis discuss and implement how to apply the feature selection technique using Rough Sets Theory combined with MLRelevance Criterion and PRelevance Criterion is. Results from feautre selection by using that method, capable of predicting the most relevant feature. So that the level of accuracy obtained able to offset, precission, recall and accuracy prior to feature selection.

Keywords : Data mining, preprocessing, classification, Rough Set, feature selection,variable selection.

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Karakteristik data pada jaman teknologi saat ini sangat bervariasi seperti pada bidang kedokteran, biologi, astronomi dan lain sebagainya yang memiliki jumlah dimensi yang tinggi [6]. Hal ini memungkinkan didapati permasalahan dengan beragamnya data dan ketidakteraturan data. Ketika data-data tersebut akan diolah untuk menghasilkan keputusan tertentu, sering kali terdapat *feature-feature* yang tidak diperlukan. Jika *feature-feature* ini tidak dihapus akan memberikan dampak negatif pada saat menganalisa data, seperti berdampak pada waktu proses yang lama, hingga yang paling buruk adalah memberikan hasil keputusan yang tidak maksimal. Hal ini memberikan sebuah permasalahan yang harus diselesaikan yaitu untuk menghilangkan *feature-feature* yang nantinya tidak akan mempengaruhi dari suatu keputusan dari data-data yang dimiliki.

Apalagi jika harus mencari data yang diinginkan pada keberagaman data tersebut, tentunya hampir tidak mungkin jika harus mencari secara manual atau satu persatu. Belum lagi jika terdapat data yang mengandung *noise, missing value, redundant* dan *irrelevant feature* [7]. Di sinilah tantangan untuk dapat mengurutkan data yang tersedia dan menemukan *feature* yang relevan untuk menjawab suatu permasalahan [12].

Di jaman teknologi informasi saat ini telah berkembang ilmu yang mempelajari tentang pemrosesan suatu data menjadi sebuah keputusan dan pengetahuan baru yaitu *Knowledge Dicovery in Database* (KDD). Di dalam KDD ini terdapat bagian dimana mampu mengatasi sebuah permasalahan dalam mengurangi atau mereduksi *feature-feature* yang tidak diperlukan atau yang sering disebut sebagai *Feature subsest selection* (FSS) atau *feature selection. Feature selection* ini mampu mengoptimalkan sebuah keputusan dimana dapat mereduksi *feature* yang sekiranya tidak diperlukan. Karena semakin kecil *feature* yang dimiliki sebuah kumpulan data, maka akan semakin maksimal hasil suatu keputusan yang didapatkan. Selain itu juga dapat menghilangkan *feature* yang redundan (berulang) [7]. Dimana setelah pada tahap *preprocessing* data akan digunakan dalam proses *data mining* yang nantinya dapat mempengaruhi hasil pengklasifikasian. Jika dalam pengklasifikasian menghasilkan hasil klasifikasi yang baik, akan mempengaruhi hasil dari sebuah keputusan [7].

Terdapat beberapa teknik dari *feature selection* yang telah muncul dalam kasus ini. *Rough Sets Theory* adalah salah satu teori matematika yang mampu menangani *feature selection* terutama dalam mereduksi *feature* dan menangani *dependency* antar *feature* [2]. *MLRelevance Criterion* merupakan sebuah formula yang dimana fokus dari perhitungannya adalah untuk memaksimalkan heterogenitas *feature* yang memiliki perbedaan kelas dan meminimalkan homogenitas *feature* yang memiliki kesamaan kelas.

Pada Tugas Akhir ini *Rough Sets Theory* akan dipadukan dengan *MLRelevance Criterion* dan *PRelevance Criterion* dimana perpaduan metode ini mampu memberikan sebuah kompleksitas sistem yang bagus [2]. Dengan perpaduan seperti ini nantinya akan mampu mengurangi kompleksitas metode Rough Set yang mempunyai komputasi sangat besar. Perpaduan metode ini juga mampu untuk meminimalkan dan mengurangi *feature* yang redundan dan mampu untuk tetap menjaga informasi penting yang terkandung di dalam data tersebut [9]. Metode *MLRelevance Criterion* bertujuan untuk mengetahui tingkat ketergantungan sebuah *feature* terhadap kelas. Metode *Rough Sets Theory* digunakan untuk mengetahui ketergantungan hubungan antar *feature* dengan kelas. Sedangkan Metode *PRelevance Criterion* digunakan untuk meminimalisir kekurangan dari metode Rough Set yang disederhanakan berkaitan dalam pembentukan jumlah kombinasi pada Tugas Akhir ini.

Hepotesa awal dari penilitan dan pengerjaan Tugas Akhir ini adalah melakukan *preprocessing* menggunakan metode *Rough Sets Theory* yang dipadukan dengan *MLRelevance Criterion* dan *PRelevance Criterion.* Metode ini diperkirakan dapat menghasilkan nilai *runing time algoritma, precision and recall,* jumlah *feature* yang diseleksi, dan waktu pembentukan model saat klasifikasi yang lebih baik. Algortima klasifikasi dipakai adalah Naive Bayes karena lebih cepat saat pemrosesan dan di dalam algoritma tersebut tidak terdapat *preprocessing* data sebelumnya.

1.2 Perumusan Masalah

Perumusan masalah dalam Tugas Akhir ini adalah yaitu:

1. Bagaimana mengimplementasikan metode *Rough Sets Theory* yang dipadukan dengan *MLRelevance Criterion* dan *PRelevance Criterion* dalam mereduksi dimensi data yang tinggi sehingga akan memudahkan dalam proses selanjutnya seperti *data mining* dan lain sebagainya.
2. Bagaimana mengukur dan menganalisis performa reduksi *data* setelah dilakukan *feature selection* dari metode *Rough Sets Theory* yang dipadukan dengan *MLRelevance Criterion* dan *PRelevance Criterion* dengan aspek parameter uji *precision, recall, accuracy* dan jumlah *feature* yang diseleksi.

1.3 Tujuan Penelitian

Adapun tujuan dalam penelitian dan pengerjaan Tugas Akhir ini adalah :

1. Menerapkan metode *Rough Sets Theory* yang dipadukan dengan *MLRelevance Criterion* dan *PRelevance Criterion* dalam proses *feature selection* untuk mereduksi dimensi data.
2 Menguraikan analisis tentang hasil pengukuran performa metode dalam *feature selection* yang diimplementasikan dengan paramater pengukuran *precision, recall, accuracy* dan jumlah *feature* yang diseleksi.

1.4 Batasan Masalah

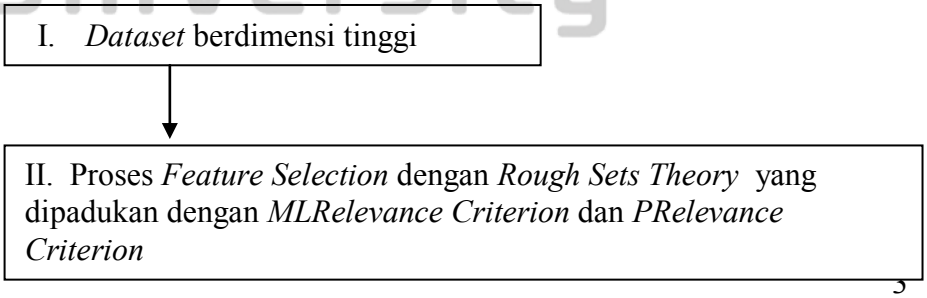
Adapun batasan masalah dalam penelitian dan pengerjaan Tugas Akhir ini adalah :

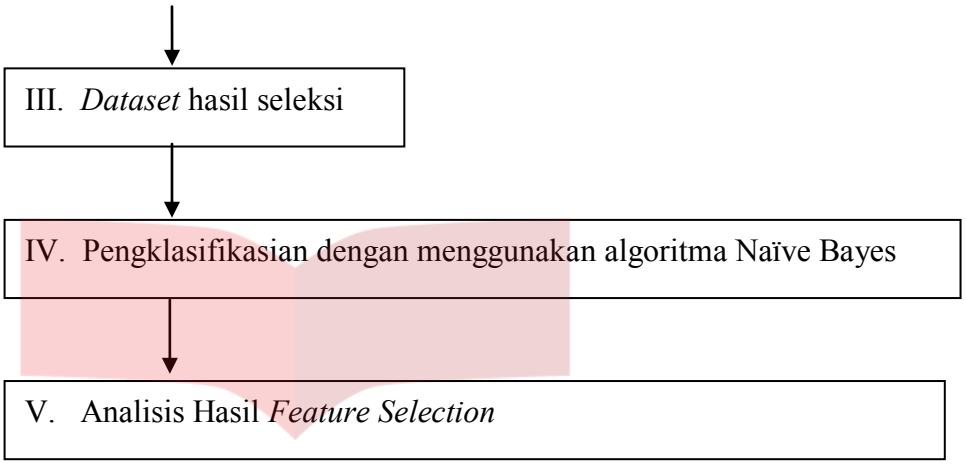
1. Penggunaan *Dataset* dengan tipe atribut diskrit yang dipublikasikan UCI Repository.
2. Tidak menangani *missing value, noisy data, data integration,* dan *data transformation.*
3. Paramater evaluasi hasil menggunakan *precision, recall, accuracy* dan jumlah *feature* yang diseleksi .
4. Algoritma klasifikasi Naives Bayes digunakan dalam evaluasi *dataset* hasil reduksi.
5. Hanya membangun aplikasi *feauture selection* yang hasilnya berupa *dataset* dengan *feature* yang telah direduksi.
6. Tidak membangun aplikasi untuk klasifikasi dengan algoritma Naïve Bayes.
7. Menggunakan aplikasi WEKA dalam melakukan pengklasifikasian *dataset* hasil *feature selection* dengan algoritma klasifikasi Naïve Bayes.

1.5 Metodologi Penyelesaian Masalah

Metode penelitian yang digunakan untuk memecahkan permasalahan dalam Tugas Akhir ini terdiri dari 5 tahap**,** yaitu :

1. Tahap Identifikasi Masalah
   Mengidentifikasi permasalahan yang ada tentang topik *data mining* khususnya *feature selection.*
2. Tahap Studi Pustaka
   Tahap ini melakukan pembelajaran, pemahaman dan pendalaman tentang materi mengenai reduksi dimensi data dan *feature selection* dengan menggunakan metode *Rough Sets Theory* yang dipadukan dengan *MLRelevance Criterion* dan *PRelevance Criterion* .
3. Tahap Pengumpulan Data
   Pada tahap ini dilakukan pengumpulan data informasi dan materi tentang *feature selection* dan metode *Rough Sets Theory* yang dipadukan dengan *MLRelevance Criterion* dan *PRelevance Criterion*
4. Tahap Desain Penelitian
   Pada tahap ini dilakukan pemodelan sistem dengan metode *Rough Sets Theory* yang dipadukan dengan *MLRelevance Criterion* dan *PRelevance Criterion* dalam *feature selection* yang mengacu pada perumusan masalah.
   Adapun penelitian awal menghasilkan alur pengerjaan sebagai berikut :





Keterangan :

I. Mengambil *Dataset* untuk dilakukan *feature selection*.

II. *Dataset* yang telah dipilih kemudian dilakukan proses *feature selection* dengan metode *Rough Sets Theory* yang dipadukan dengan *MLRelevance Criterion* dan *PRelevance Criterion* untuk menghasilkan *feature* yang dirasa penting, menghilangkan *feature* yang redundan serta, meminilisir homogenitas.

III. Dari proses ke dua, dihasilkan sebuah *dataset* yang telah dilakukan *feature selection*.

IV. Dari hasil *feature selection* tadi, kemudian *dataset* diproses klasifikasinya dengan menggunakan naives bayes.

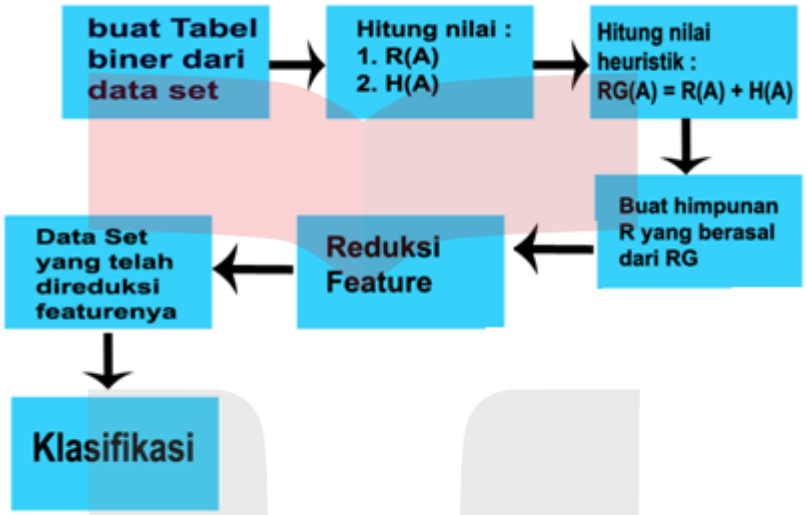
V. Dari hasil klasifikasi tersebut akan dibandingkan dengan hasil klasifikasi yang tanpa melalui proses preprocessing dengan *feature selection* dengan perbandingan *precision*, *recall*.

VI. Dalam pelaksanaan nantinya, apkliasi ini akan menggunakan Matlab dalam pengimplementasian metode *Rough Sets Theory* yang dipadukan dengan *MLRelevance Criterion* dan *PRelevance Criterion.* Kemudian untuk klasifikasinya dilakukan dengan bantuan tool Weka yang memakai algoritma Naives Bayes.

5. Tahap Implementasi
   Dari pemodelan sistem yang telah dibuat kemudian diimplementasikan menjadi sebuah sistem reduksi data dengan metode *Rough Sets Theory* yang dipadukan dengan *MLRelevance Criterion* dan *PRelevance Criterion* dengan bantuan tool Matlab.
6. Tahap Pengujian dan Analisis
   Tahap ini melakukan pengujian dan analisi untuk mengetahui apakah tujuan dari pembuatan dan penilitian Tugas Akhir ini telah tercapai.
7. Tahap Pembuatan Laporan

Tahap ini adalah tahap terakhir dimana dilakukan pembuatan laporan dan pendokumentasian dari sistem reduksi data ini.

Deskripsi Sistem
sebagai berikut :



Gambar 1 : Gambaran umum proses sistem

1. Membuat tabel biner dari *dataset* yang dipilih. Baris menyatakan pasangan objek, dan kolom menyatakan tiap atribut. Ditambah kolom terakhir berupa *decission* yang dianggap sebagai atribut.
2. Kemudian setiap *feature* dihitung nilai R(A) dan H(A). Metode Rough Sets berperan dalam menentukan nilai H(A).
3. Untuk atribut 'A' dihitung nilai RP(A) dan bentuk sebuah list atribut secara berurutan mulai dari atribut yang paling relevan (nilai RP(A) yang paling besar)
4. Melakukan reduksi data dengan nilai *Threshold* yang dibandingkan dengan nilai RP(A)
5. Didapat *dataset* yang telah direduksi atribut yang dinilai tidak perlu.

1.6 Sistematika Penulisan

Penulisan tugas akhir ini dibagi ke dalam 5 bab, antara lain :

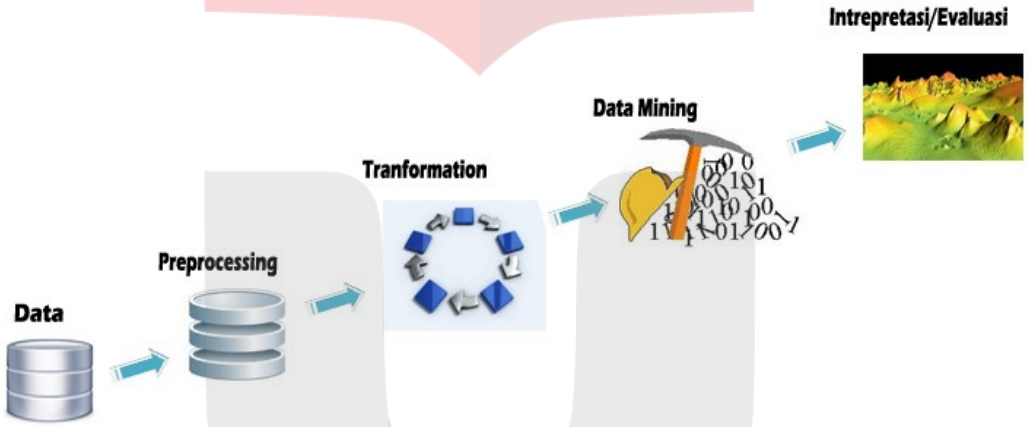
1. **Bab 1 Pendahuluan**
   Berisi latar belakang masalah, perumusan masalah, batasan masalah, tujuan masalah, metodologi penyelesaian masalah dan sistematika penulisan

2. **Bab 2 Landasan Teori**
   Berisi landasan teori tentang KDD, metode *MLRelevance Criterion* , *PRelevance Criterion, Rough Sets Theory* dan klasifikasi.
3. **Bab 3 Perancangan Aplikasi**
   Berisi analisa dan perancangan aplikasi yang akan dibuat dengan bahasa permodelan DFD
4. **Bab 4 Analisa Hasil**
   Berisi penjelasan mengenai implementasi hasil perancangan, uji coba terhadap sistem, dan analisa perangkat lunak yang dibangun.
5. **Bab 5 Kesimpulan dan Saran**
   Berisi kesimpulan dan saran dari hasil penelitian untuk pengembangan lebih lanjut.

BAB II
LANDASAN TEORI

2.1 *Knowledge Discovery in Databases (KDD)*

Knowledge Discovery in Database yang selanjutnya disebut KDD adalah suatu keseluruhan proses konversi data mentah menjadi suatu pengetahuan yang bermanfaat. Proses KDD meliputi serangkaian tahap transformasi meliputi *data preprocessing* dan *post-processing.* Proses-proses yang terdapat dalam *KDD* yaitu *[3] :*



Gambar 2 : Proses dalam KDD

1. Data *Selection*

   Melakukan pemilihan data pada kumpulan data yang akan digunakan sebelum masuk ke dalam tahap penggalian informasi dalam KDD dimulai. Kemudian data hasil dari pemilihan tersebut di simpan dalam suatu berkas terpisah dari data asalnya sebelum dilakukan proses selanjutnya.

2. *Preprocessing*

   Data dari hasil pemilihan pada tahap sebelumnya, kemudian dilakukan proses *cleaning* agar data efisien saat dilakukan proses di *data mining*.

3. *Transformation*

   Coding adalah proses transformasi pada data yang telah dipilih sehingga data tersebut sesuai untuk proses *data mining.* Proses *coding* sangat fleksibel dan tergantung pada jenis atau pola informasi yang akan dicari pada basis data.

4. *Data mining*

   Data mining adalah proses ekstraksi informasi atau pola yang penting atau menarik dari data yang ada di database yang besar. Proses *data mining* meliputi klasifikasi, klastering, asosiasi dll.

5. Intrepretation / Evaluation

BAB V
Kesimpulan dan Saran

5.1 Kesimpulan

Dari hasil penilitan diatas, dapat diambil kesimpulan :

1. Proses *feature selection* menggunakan perpaduan metode *Rough Sets, MLRelevance Criterion* dan *PRelevance Criterion* dapat dikatakan berhasil, karena dapat mengimbangi nilai *precision*, *recall* dan *accuracy* dari *dataset* awal.
2. Dengan menentukan nilai P=2 dalam pembentukan kombinasi dapat menghasilkan proses *feature selection* yang relatif baik jika melihat perbandingan nilai *precision*, *recall* dan *accuracy* dari *dataset* awal dengan *dataset* hasil *feature selection* serta mampu melakukan proses komputasi lebih cepat dibandingkan dengan P>2.
3. Nilai *threshold* dapat digunakan dalam proses *feature selection*. Semakin besar nilai *threshold* maka semakin banyak jumlah *feature* yang dapat dihilangkan.
4. Dengan menggunakan nilai *threshold* lebih menguntungkan karena dapat memilih jumlah *feature* yang diinginkan untuk dihilangkan tanpa mengurangi nilai *precision*, *recall*, dan *accuracy*, bahkan untuk meningkatkannya.
5. *Feature selection* dapat meningkatkan nilai performansi *precision*, *recall* dan *accuracy* jika tepat dalam melakukan pemilihan *feature*.
6. Karakteristik *dataset* mempengahuri hasil dari *feature selection*.
7. *Feature selection* akan lebih optimal jika menggunakan *dataset* yang mempunyai jumlah *feature* sedang hingga banyak.

5.2 Saran

1. Melakukan percobaan terhadap *dataset* yang dengan jumlah *feature*, objek, dan variasi jenis nilai dalam setiap *feature* yang lebih banyak dan lebih bervariasi jumlahnya serta menggunakan data yang mengandung nilai numerik dimana terdapat nilai real yang kemudian dilakukan proses diskritisasi agar menjadi data yang bersifat kategorikal.

Daftar Pustaka

[1] Adiwijaya.. *Diktat Matematika Diskrit*. IT Telkom. Bandung

[2] Caballero, Yailé, Rafael Bello, Delia Alvarez, Maria M. Garcia. 2006. *Two new feature selection algorithms with Rough Sets Theory*. [online]. (http://pitagoras.usach.cl/~gfelipe/wcc/papers/AI/Article_22-Caballero.pdf, diakses tanggal 19 Oktober 2010).

[3] Firestone, Joseph M. 1997. *Data mining and KDD: A Shifting Mosaic*. [online], (http://tinyurl.com/42j4pr4, diakses tanggal 11 Febuari 2010)

[4] Greco , Salvatore, Benedetto Matarazzo, Roman Slowinski. 1999. *Rough Sets Theory For Multicriteria Decision Analysi*. [online]. (http://tinyurl.com/3jbef2l, diakses tangal 19 Oktober 2010)

[5] Hall, Mark A. 1999. *Correlation-based Feature selection for Machine Learning*.[online]. (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.9584&rep=rep1&type=pdf, diakses pada tanggal 19 Oktober 2010)

[6] Han, Jiawei. Kamber, Micheline. 2001. *Data Mining Concepts and Techniques.* San Francisco: Morgan Kaufmann Publisher. Inc

[7] Karima, Putri Hapsari. 2010. *Implementasi dan Analisis Feature selection Menggunakan Mutual Information & Redundancy-Synergy Coefficient.* IT Telkom. Bandung

[8] Neil S. Mac Parthal´ain. 2009.*Rough Set Extensions for Feature selection*. Department of Computer Science Aberystwyth University.

[9] Piñero, P; Arco, L; García, M. and Caballero, Y. *Two New Metrics for Feature selection in Pattern Recognition*. [Online]. (http://www.springerlink.com/content/f5h8jcv9h0cwdg85/, Diakses tanggal 21 Oktober 2010.

[10] Puspitarani, Yan. 2008. *Filter-Based Feature selection pada Kategorisasi Artikel Berita Berbahasa Indonesia.* IT Telkom. Bandung.

[11] Sebban, Marc. 1999. *On Feature selection: a New Filter Model.* [Online]. (http://www.aaai.org/Papers/FLAIRS/1999/FLAIRS99-041.pdf, diakses tanggal 19 Oktober 2010)

[12] Swiniarki. R.W. and Skowron. A. *Rough Set Methods in Feature selection and recognition.* [Online], (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.106.354&rep=rep1&type=pdf, diakses tanggal 19 November 2010)

[13] Ambarita, Ardedi Frianto. 2008. *Penggunaan Rough Set Approach Sebagai Kriteria Variable Selection Dalam Task Classification Pada Data Mining* . IT Telkom. Bandung