Statistika, Vol. 11 No. 1, 61 – 66
Mei 2011

Feature Selection Data Indeks Kesehatan Masyarakat Menggunakan Algoritma *Relief*

ZURNILA MARLI KESUMA

Program Studi Matematika FMIPA Universitas Syiah Kuala Banda Aceh
Jl. Syech Abdurrauf no. 2 Kopelma Darussalam Banda Aceh
email: bukulily@gmail.com

ABSTRAK

Feature selection adalah suatu metode penganalisaan data yang bertujuan untuk memilih fitur yang berpengaruh (fitur optimal) dan mengesampingkan fitur yang tidak berpengaruh. Ada beberapa algoritma *feature selection* yang dapat digunakan, salah satunya adalah *Relief*. *Relief* memanfaatkan teknik bobot (*weight*) untuk mengukur signifikansi fitur dalam konteks klasifikasi dan fitur yang memiliki nilai bobot di atas ambang batas (*threshold*) yang digunakan akan dipilih. Penelitian ini bertujuan untuk mendapatkan fitur optimal dari data data indeks kesehatan masyarakat.
Hasil pengolahan data menunjukkan bahwa untuk setiap data yang diuji hanya menghasilkan satu fitur optimal dengan nilai *threshold* yang berbeda.
Kata Kunci: feature selection, algoritma relief, threshold, weight dan fitur optimal

1. PENDAHULUAN

Suatu objek perlu diketahui fitur-fiturnya agar dapat dikenali dan dibedakan dari objek yang lain. Fitur-fitur optimal yang dapat diketahui dari suatu objek akan mempermudah dan mempercepat proses identifikasi objek tersebut. Menurut Sugiyono (1997) dalam Umar (1998), fitur atau variabel di dalam penelitian merupakan suatu atribut dari sekelompok objek yang diteliti yang mempunyai variasi antara satu dengan yang lain dalam kelompok tersebut. Sedangkan *Feature Selection* adalah suatu kegiatan pemodelan atau penganalisaan data yang umumnya dapat dilakukan secara *preprocessing* dan bertujuan untuk memilih fitur yang berpengaruh (fitur optimal) dan mengesampingkan fitur yang tidak berpengaruh, (Rehat, 2009). Ada beberapa algoritma *Feature Selection* yang dapat digunakan. untuk menemukan fitur-fitur yang optimal dari sebuah himpunan fitur. Salah satu algoritma *Feature Selection* adalah algoritma *Relief*. *Relief* pertama kali diusulkan oleh Kira dan Rendell pada tahun 1992. *Relief* termasuk dalam metode *Feature Selection* tipe *Filter*, yang didasarkan pada estimasi fitur. *Relief* memberikan nilai yang relevan untuk setiap fitur, dan fitur yang memiliki nilai di atas ambang batas (*threshold)* yang diberikan oleh pengguna yang akan dipilih. Algoritma *Relief* memanfaatkan teknik bobot untuk mengukur signifikansi fitur dalam konteks klasifikasi. Bobot *Relief* adalah nilai-nilai yang kontinu dan memungkinkan fitur untuk digolongkan berdasarkan relevansi. *Relief* juga merupakan algoritma yang menarik dalam *Feature Selection* karena memiliki komputasi yang efisien, (Chouchoulas, 2001).

2. FEATURE SELECTION

Feature Selection adalah suatu proses yang mencoba untuk menemukan subhimpunan dari himpunan fitur yang tersedia untuk meningkatkan aplikasi dari suatu algoritma pembelajaran. *Feature Selection* digunakan dibanyak area aplikasi sebagai alat untuk menghilangkan fitur yang tidak relevan dan atau fitur berlebihan. Sebuah fitur dikatakan tidak relevan jika memberikan sedikit informasi, sedangkan sebuah fitur dikatakan berlebihan jika informasi yang diberikan adalah informasi yang terkandung dalam fitur lain (tidak memberikan informasi baru).

Ada empat langkah yang dilakukan dalam feature selection *(Dash, 1997)* yaitu:

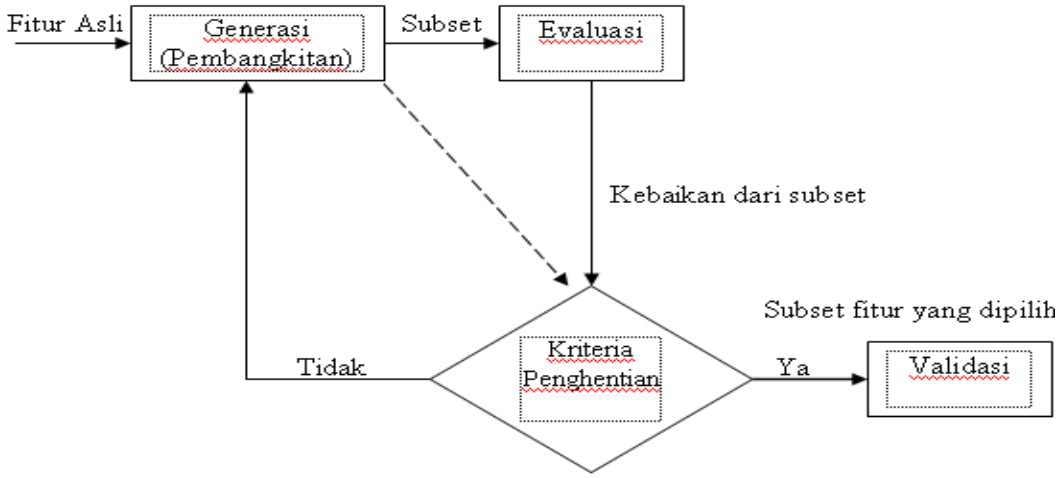
1.Prosedur generasi (pembangkitan), untuk menghasilkan calon subhimpunan berikutnya dapat dilakukan dengan beberapa cara yaitu : lengkap, heuristik dan acak.

2. Evaluasi fungsi, untuk mengevaluasi subhimpunan, dengan cara mengukur jarak, informasi, konsistensi, ketergantungan, dan mengukur tingkat kesalahan klasifikasi.

3. Kriteria penghentian, untuk memutuskan kapan harus berhenti, dengan cara melihat nilai ambang batas *(threshold)*, diawali dengan sejumlah pengulangan dan sebuah ukuran subhimpunan fitur terbaik.

4. Prosedur validasi, untuk memeriksa apakah subhimpunan valid. (opsional).

Proses dalam feature selection tersebut dapat dituangkan dalam skema berikut:



Gambar 1. Proses *Feature Selection* dengan validasi, (Dash dan Liu, 1997)

2.1. Prosedur Generasi

Prosedur generasi merupakan prosedur pencarian yang pada dasarnya menghasilkan *subset* (subhimpunan) dari fitur-fitur untuk dievaluasi. Jika himpunan fitur asli berisi *N* jumlah fitur, maka jumlah calon bersaing untuk menjadi subhimpunan yang dihasilkan adalah *2N*. Ini merupakan jumlah besar bahkan untuk setengah dari jumlah *N*. Ada berbagai pendekatan untuk menyelesaikan masalah ini, yaitu: lengkap, heuristik, dan acak.

(a) Lengkap

Urutan ruang pencarian prosedur generasi ini adalah *O*(*2N*), sebuah subhimpunan yang sedikit untuk dievaluasi. Subhimpunan fitur yang optimal sesuai dengan evaluasi fungsi, karena prosedur ini dapat dilakukan dengan cara mundur. Mundur dapat dilakukan dengan menggunakan berbagai teknik, seperti: *branch and bound*, pencarian pertama terbaik, dan balok pencarian.

(b) Heuristik

Dalam setiap pengulangan prosedur generasi ini, semua sisa fitur yang belum dipilih (ditolak) masih dipertimbangkan untuk pemilihan (penolakan). Ada banyak variasi untuk proses sederhana ini, tapi generasi subhimpunan pada dasarnya meningkat atau menurun. Urutan ruang pencarian adalah *O*(*N2*) atau kurang. Prosedur ini sangat sederhana untuk diterapkan dan sangat cepat dalam memperoleh hasil, karena ruang pencarian hanya kuadrat dari jumlah fitur.

(c) Acak

Prosedur generasi ini masih baru dalam penggunaannya dalam metode *Feature Selection* dibandingkan dengan dua kategori lainnya. Meskipun ruang pencarian adalah *O*(*2N*), tetapi metode ini biasanya mencari lebih sedikit jumlah subhimpunan daripada *2N* dengan menetapkan jumlah maksimum pengulangan yang mungkin. Optimalitas subhimpunan yang dipilih tergantung pada sumber daya yang tersedia. Setiap prosedur generasi acak akan memerlukan nilai-nilai dari beberapa parameter.

2.2. Evaluasi Fungsi

Evaluasi fungsi mengukur kebaikan subhimpunan yang dihasilkan oleh beberapa prosedur generasi, dan nilai ini dibandingkan dengan yang terbaik sebelumnya. Jika ditemukan yang lebih baik, maka subhimpunan terbaik sebelumnya digantikan. Ada beberapa cara dalam melakukan evalusi fungsi, salah satunya yaitu ukuran Jarak**.**

Juga dikenal sebagai keterpisahan, perbedaan, atau diskriminasi ukuran. Untuk dua kelas, fitur *X* adalah fitur yang lebih disukai dari fitur *Y* apabila *X* menginduksi perbedaan yang lebih besar antara kedua kelas probabilitas kondisional dari *Y* dan jika perbedaan adalah nol, maka *X* dan *Y* tidak dapat dibedakan (sama). Sebagai contoh adalah jarak *Euclidea*. *Euclidean* merupakan metode pengukuran jarak di antara dua objek berdasarkan akar jumlah kuadrat jarak kedua objek. Rumus umum untuk menghitung jarak *Euclidean* yaitu, jika *X* memiliki koordinat ($x_1, x_{2,...}x_n$) dan objek *Y* memiliki koordinat ($y_1, y_2, ...y_n$), maka jarak *Euclidean* kedua objek tersebut adalah, $d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)}$

2.3. Kriteria Penghentian

Prosedur generasi dan evaluasi fungsi dapat mempengaruhi pilihan untuk kriteria penghentian.

2.4. Prosedur Validasi

Proses validasi bukan merupakan bagian dari proses *Feature Selection* itu sendiri, namun sebuah *Feature Selection* harus divalidasi dengan cara melakukan pengulangan terhadap evaluasi fungsi subhimpunan dari fitur sampai kriteria penghentian terpenuhi, (Dash, 1997).

3. ALGORITMA *RELIEF*

Dalam Kira (1992), *Algoritma relief secara umum sebagai berikut:*

Relief (δ, m, τ)

Separate δ*into* δ^+ *(positive instances) and* δ^- *(negative instances).*
W=(0,0,. . . , 0)
For i=1 to m
Pick at random an instance $X \varepsilon \delta$
Pick at random one of the positive instances
Closest to $X, Z^+ \varepsilon\ \delta^+$
Pick at random one of the negative instances
Closest to $X, Z^- \varepsilon\ \delta^-$
If (X is a positive instance)
Then Near-hit=Z^+*; Near-miss =* Z^-
Else Near-hit=Z^-*; Near-miss =* Z^+
Update-weight (W,X,Near-hit,Near-miss)
Relevance=(1/m)W
For i= 1 to p

If (relevance ≥τ)
Then f_i is a relevant feature
Else f_i is a irrelevant feature
Update-weight (W,X,Near-hit,Near-miss)

For i= 1 to p
$Wi= Wi – diff(xi, near\text{-}hit_i)^2 + diff(xi, near\text{-}miss_i)^2$

Figure 1. Relief Algoritm

4. THRESHOLD T (AMBANG BATAS)

Threshold (ambang batas) merupakan nilai batas relevan untuk pemilihan fitur optimal. Nilai *threshold* berada pada interval 0 sampai 1 dan penggunaannya bersifat independen (tergantung pada pengguna). Dalam algoritma *Relief, threshold* akan dibandingkan dengan nilai *weight* (bobot) dari suatu fitur. Apabila suatu fitur memiliki nilai bobot lebih besar dari *threshold* yang digunakan maka fitur tersebut merupakan fitur optimal sedangkan jika nilai bobot fitur lebih kecil sama dengan dari *threshold* maka fitur tersebut tidak akan dipilih karena tidak termasuk dalam kategori fitur optimal, (Kira, 1992).

5. IMPLEMENTASI

Data yang digunakan dalam penelitian ini adalah data sekunder dari hasil Indeks Pembangunan Kesehatan Masyarakat (IPKM). Dan definisi IPKM adalah indikator komposit yang menggambarkan kemajuan pembangunan kesehatan. IPKM merupakan indeks komposit yang dirumuskan dari 24 indikator kesehatan (sumber data kesehatan utama Riskesdas 2007), dan berdasarkan indikator komposit tersebut dibuat peringkat kab/kota, dari peringkat terbaik sampai ke peringkat terbawah.

5.1. Identifikasi Fitur (Variabel)

IPKM disusun dengan tujuan untuk menterjemahkan acuan pembangunan daerah saat ini khususnya untuk pembangunan di bidang kesehatan dan daerah bisa melakukan penajaman program intervensi di bidang kesehatan dengan berdasarkan variable yang terangkum dalam indeks tersebut. IPKM tersusun dari berbagai indikator kesehatan yang berdasarkan kajian secara mendalam bersama para pakar kesehatan baik pakar kesehatan yang berada di dalam institusi kementerian kesehatan maupun dari berbagai perguruan tinggi terpilih 24 variabel. Ke 24 fitur yang telah terpilih tersebut terdiri dari:

1. prevalensi balita gizi buruk dan kurang
2. prevalensi balita sangat pendek dan pendek,
3. prevalensi balita sangat kurus dan kurus,
4. prevalensi balita gemuk,
5. prevalensi diare,
6. prevalensi pnemonia,
7. prevalensi hipertensi,
8. prevalensi gangguan mental,
9. prevalensi asma,
10. prevalensi penyakit gigi dan mulut,
11. prevalensi disabilitas,
12. prevalensi cedera,
13. prevalensi penyakit sendi,
14. prevalensi ISPA,
15. proporsi perilaku cuci tangan,
16. proporsi merokok tiap hari,
17. akses air bersih,
18. akses sanitasi,

19. cakupan persalinan oleh nakes,
20. cakupan pemeriksaan neonatal-1
21. cakupan imunisasi lengkap,
22. cakupan penimbangan balita,
23. ratio Dokter/Puskesmas, dan
24. ratio bidan/desa.

5.2. Prosedur

Prosedur yang dilakukan adalah sebagai berikut :

1. Membagi label data menjadi *binary classification* yaitu berupa kelas positif dan kelas negatif.
2. Menentukan *threshold* yang akan digunakan. Dalam penelitian ini *threshold* yang akan digunakan adalah 0.01, 0.02, 0.04 dan 0.06.
3. Memanggil (*import*) data dari program *Microsof Excel* dengan menggunakan perangkat lunak R.2.12.1.
4. Langkah 2 dan 3 diulang sebanyak 10 kali untuk setiap *threshold* yang digunakan dari masing-masing data. Dari banyaknya pengulangan yang dilakukan, dihitung berapa nilai *weight* (bobot) dari tiap fitur.
5. Membandingkan hasil nilai *weight* (bobot) dari tiap fitur tersebut dengan nilai *threshold* yang digunakan dan mendapatkan simpulan fitur mana yang terbaik (optimal). Suatu fitur dikatakan optimal apabila nilai bobotnya lebih besar dari nilai *threshold* yang digunakan.

Tabel 1. Analisa data untuk threshold=0.01, 0.02 dan 0.04

Threshold = 0.01

No	Feature	Freq.	Weight
[1,]	19	10	0.057975
[2,]	22	10	0.048738
[3,]	12	10	0.045632
[4,]	23	10	0.044766
[5,]	21	10	0.039404
[6,]	15	10	0.037991
[7,]	10	10	0.032633
[8,]	7	10	0.03118
[9,]	16	10	0.028602
[10,]	4	10	0.028362
[11,]	18	10	0.02594
[12,]	20	10	0.025026
[13,]	2	10	0.02406
[14,]	6	10	0.024027
[15,]	11	10	0.021837
[16,]	13	10	0.020355
[17,]	1	9	0.020306
[18,]	17	10	0.019932
[19,]	8	10	0.019368
[20,]	9	9	0.017347
[21,]	3	9	0.014441

Threshold = 0.02

No	Feature	Freq	Weight
[1,]	19	100	0.06092
[2,]	23	100	0.047098
[3,]	22	100	0.043868
[4,]	12	100	0.043186
[5,]	21	100	0.039419
[6,]	15	100	0.0366
[7,]	10	90	0.028712
[8,]	18	90	0.027462
[9,]	16	90	0.025484
[10,]	7	70	0.025114
[11,]	6	70	0.023857
[12,]	4	70	0.023661
[13,]	20	60	0.023384
[14,]	2	80	0.023176
[15,]	1	60	0.02114
[16,]	17	50	0.020218
[17,]	8	50	0.020107

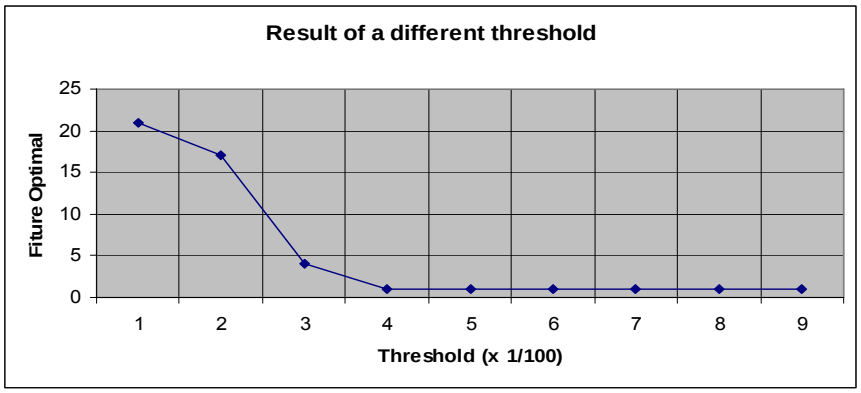
Threshold = 0.04

No	Feature	Freq	Weight
[1,]	19	10	0.057869
[2]	22	6	0.048738
[3]	12	9	0.045632
[4]	23	7	0.044766

5.3 Pengujian data IPKM dan Hasil

Dari hasil analisa algoritma *Relief* diperoleh feature sebagaimana yang ada pada Tabel 1. Gambar 1 menunjukkan terjadinya penurunan ekstrim di threshold 0,01 ke 0,02, dari 21 feature menjadi hanya 17 fitur yang optimal. Untuk mendapatkan nilai yang optimal, masih perlu dilakukan langkah validasi.

Berdasarkan hasil yang diperoleh, untuk threshold 0.02 , 17 fitur yang terpilih adalah: 19, 23, 22, 12, 21, 15, 10, 18, 16, 7, 6, 4, 20, 2, 1, 17, 8.



Gambar 1. Grafik hasil threshold

DAFTAR PUSTAKA

[1]. Acuna, Edgar, and members of the CASTLE group at UPR-Mayaguez and Puerto Rico, dprep: Data preprocessing and visualization functions for classification, R package version 2.0 (2008).
[2]. Chouchoulas, A., Incremental Feature Selection Based On Rough Set Theory PhD Proposal Centre for Intelligent Systems and Applications Division of Informatics , The University of Edinburgh, Scotland (2001).
[3]. Dash, M. and H. Liu, Feature Selection for Classification, Intelligent Data Analysis,1(1-4) : 131-156 (1997).
[4]. Guyon, I and A. Elisseeff, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3, 1157-1182 (2003).
[5]. Kira, K. and L. A. Rendell, A Practical Approach to Feature Selection, In Proceedings of the Ninth International Workshop on Machine Learning, Morgan Kaufmann Publishers Inc., 249-256 (1992a).
[6]. Kira, K. and L. A. Rendell, The Feature Selection Problem : Traditional Methods and a New Algorithm, In Procedding of Tenth National Conference on Artificial Intelligence, MIT Press, 129-134 (1992b).

,