ANALISIS DAN IMPLEMENTASI ALGORITMA RELIEFF UNTUK FEATURE SELECTION PADA KLASIFIKASI DATASET MULTICLASS

Danang Aji Irawan[1], Z.k. Abdurahman Baizal[2], Erda Guslinar Perdana[3]

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

Abstrak

Permasalahan yang terjadi pada data adalah jumlah yang terlalu banyak. Salah satunya adalah pada jumlah atribut yang ada dalam data tersebut. Untuk menanganinya , kita perlu melakukan reduksi data pada dimensi atribut. Teknik ini biasanya disebut dengan Feature Selection atau pemilihan atribut. Teknik ini merupakan salah satu dari teknik yang dilakukan pada data preprocessing. Tujuan melakukan feature selection ini, selain untuk mereduksi jumlah atribut, nantinya bisa memberikan performansi yang lebih baik pada saat melakukan klasifikasi dibandingkan menggunakan data yang tidak dilakukan pemilihan atribut.

Pada Tugas Akhir ini, penulis mengimplementasikan salah satu algoritma feature selection yaitu ReliefF. Algortima ini merupakan algoritma pemilihan atribut yang berbasis pada instan atau record. Pemilihan atribut dilakukan dengan menghitung perbedaan bobot untuk tiap instan yang terpilih secara acak (random sampling) dengan instan yang terpilih sebagai near hit (tetangga terdekat instan terpilih pada kelas yang sama) dan near miss (tetangga terdekat instan terpilih pada kelas yang berbeda).

Tahap penghitungan performansi akan didasarkan pada data baru yang berisi data dengan atribut yang terpilih untuk kemudian dilakukan proses klasifikasi. Dari proses klasifikasi ini, akan dihitung perbedaan performansi dari data yang belum dilakukan pemilihan fitur dengan yang telah dilakukan proses pemilihan fitur. Hasil yang ditangkap adalah nilai dari precission dan recall.

Hasil implementasi, pengujian, dan analisis pada Tugas Akhir ini menunjukkan bahwa kinerja algoritma ini sangat bergantung pada jumlah iterasi yang dilakukan, jumlah tetangga terdekat, penentuan threshold, dan juga kualitas instan yang terpilih secara acak saat algoritma ini dijalankan. Dataset yang telah mengalami proses pemilihan fitur telah mampu meningkatkan performansi hasil klasifikasi.

Kata Kunci : feature selection, ReliefF, klasifikasi, near hit, near miss

Abstract

Problems that occurred in the data is the amount that is too big. One of the problem is the number of attributes that exist in the data. To handle the problem, we need to make data reduction on dimension of attribute. This technique is called Feature Selection. This technique is one of the techniques performed on the data preprocessing. The purpose of doing feature selection, in addition to reducing the number of attributes, will be able to give a better performance of classification when we compared to data that without attribute selection.

In this research, the author implements a feature selection algorithm called ReliefF. This algorithm is attribute selection algorithm that based in instance. The selection attributes is done by calculating differences weight for each instance with selected by randomly as near hit (nearest neighbor was elected in the same class) and near miss (nearest neighbor was elected in the different class).

Calculation of performance will be based on new data that contains data with attribute that are selected from feature selection for the classification. From the classification process, difference of performance will be calculated the data that has not been done with the selection of features that have made the process of selection of features. Results are captured the value of precision and recall.

The results of the implementation, testing, and analysis on this Final Project shows that the performance of the algorithm is highly dependent on the number of iterations performed, the number of nearest neighbors, election of threshold and also the quality of the randomly selected instant when the algorithm is run. Dataset that has undergone a process of selection of features has been able to improve the performance of the classification results.

Keywords : feature selection, ReliefF, classification, near hit, near miss.

Bab 1 Pendahuluan

1.1. Latar Belakang Masalah

Sebuah data yang ada saat ini bisa memiliki dimensi yang sangat besar, baik dari banyaknya instan ataupun atribut yang dimiliki. Misalnya saja untuk data kependudukan, satu orang bisa memiliki atribut yang banyak. Orang bisa memiliki data tentang nama, alamat, nomor telepon, usia, gaji, dan lain sebagainya. Banyaknya atribut yang bisa dimiliki oleh satu objek belum tentu merupakan informasi relevan yang dibutuhkan oleh sistem data mining. Untuk itulah perlu dilakukan proses reduksi data. *Feature Selection* merupakan cara yang efektif untuk melakukan reduksi data dan menjadi langkah penting yang perlu dilakukan supaya aplikasi data mining berhasil dengan baik[4].

Feature selection adalah suatu proses memilih subset dari fitur/atribut yang optimal dengan menggunakan kriteria tertentu. *Feature selection* merupakan salah satu dari proses *pre-processing* pada suatu *dataset* yang akan dilakukan proses data mining. Dengan melakukan *feature selection* ini mampu untuk mengurangi jumlah *feature* yang tidak relevan, menghilangkan redundan data, menghilangkan *feature* yang mengandung *noisy* dan akan memberikan efek meningkatkan kecepatan dalam melakukan data mining, meningkatkan akurasi *learning*, dan menghasilkan model yang baik[4].

Telah banyak teknik-teknik *feature selection* yang berkembang hingga saat ini. ReliefF yang ditemukan oleh Kononenko pada 1994 sebagai salah satu teknik *feature selection*, merupakan pengembangan dari algoritma Relief yang dikembangkan pada tahun 1992 oleh Kira dan Rendell [1,2]. Relief memberikankan estimasi atribut yang sangat efisien. Ide dasarnya adalah menghitung nilai perbedaan jarak antar instan. Perbedaan jarak ini nantinya akan digunakan untuk melakukan penghitungan bobot. Ukuran jarak yang digunakan pada algoritma Relief ada 2 macam yaitu *nearhit* (jarak antar instan dalam satu kelas) dan *nearmiss*(jarak antar instan yang berbeda kelas).

Relief sendiri pada dasarnya bisa digunakan pada atribut diskrit dan kontinu tapi memiliki keterbatasan untuk melakukan seleksi fitur pada *dataset* 2 kelas saja. ReliefF mampu menyempurnakan kelemahan algoritma Relief tersebut. Penyempurnaan yang dilakukan oleh Kononenko pada algoritma ReliefF yaitu bisa menangani *dataset* yang *multiclass*, data tidak lengkap (*incomplete data*) dan data kotor (*noisy data*) [3]. ReliefF termasuk dalam algoritma yang menggunakan fungsi evaluasi *distance measure* sehingga memiliki waktu kompleksitas yang rendah[8]. Penanganan pada *dataset* dengan data yang besar, jumlah *instance* lebih dari 300, bisa kita lakukan dengan menggunakan *sampling* secara acak dari *dataset* tersebut[3].

Untuk itulah pada Tugas Akhir ini, penulis mencoba menggunakan ReliefF untuk melakukan *feature selection* pada data tersebut sesuai dengan kemampuan yang dimiliki algoritma ini.

1.2. Perumusan Masalah

Berdasarkan pada latar belakang diatas permasalah yang difokuskan pada Tugas Akhir ini adalah

1. Bagaimana implementasi algoritma ReliefF untuk melakukan proses *feature selection* pada *dataset multiclass*.

2. Bagaimana pengaruh pemilihan jumlah tetangga terdekat , jumlah sampel instan , dan pemilihan *threshold* terhadap performansi algoritma ReliefF serta jumlah fitur yang dipilih.
3. Bagaimana performansi pada data yang sudah dilakukan proses *feature selection* menggunakan algoritma ReliefF berdasarkan nilai *precission* dan *recall* yang dihasilkan oleh *classifier*.

Dalam implementasi tugas akhir ini akan dibatasi oleh beberapa hal yaitu:

1. *Dataset* yang akan digunakan berasal dari UCI Machine Learning Repository, dan KEEL *dataset repository* dengan memperhatikan bahwa *dataset* tersebut merupakan *dataset multiclass*.
2. *Dataset* yang akan dipakai memperhatikan jumlah instan lebih dari 800 instan.
3. Proses klasifikasi akan dilakukan menggunakan metode *classifier* IB1 yang terdapat pada Weka 3.6.3. Hal ini karena IB1 adalah salah satu *classifier* berbasis pada instan dan menggunakan *k-nearest neighbor*. Memiliki karakteristik yang hampir sama dengan algoritma *feature selection* ReliefF yang akan digunakan pada Tugas Akhir ini. Selain itu, digunakan *classifier* Naïve Bayes sebagai *classifier* lain yang akan digunakan untuk menguji performansi ReliefF, yang tidak berhubungan dengan penggunaan *k-nearest neighbor* di dalamnya.

1.3. Tujuan

Tujuan dari Tugas Akhir ini adalah :

1. Mengimplementasikan algoritma ReliefF untuk melakukan proses *feature selection*.
2. Menganalisis pengaruh pemilihan jumlah tetangga terdekat dan jumlah sampel instan dalam proses *feature selection* menggunakan algoritma ReliefF.
3. Menganalisis performansi pengukuran *classifier* yang meliputi *precision* dan *recall* pada *dataset* yang belum dan sudah dilakukan proses *feature selection*.

1.4. Metodologi Penyelesaian Masalah

Metodologi yang akan dilakukan adalah :

1. Studi literatur

Pada tahap ini akan dilakukan untuk mempelajari konsep dan teori-teori pendukung yang akan digunakan untuk melakukan seleksi fitur. Pembelajaran yang perlu dilakukan meliputi :

1. Konsep dasar tentang *feature selection*.
2. Konsep dasar melakukan *feature selection* menggunakan algoritma ReliefF.
3. Konsep dasar mengenai pencarian tetangga terdekat suatu instan pada *dataset* numerik dan diskrit (kategori dan nominal).
4. Konsep dasar untuk penghitungan performansi dari nilai *precission* dan *recall*.

Serta dengan mencari informasi-informasi lain yang bisa menunjang Tugas Akhir ini.

2. **Pengumpulan data**

   Tahap ini akan dilakukan pencarian data-data pendukung terutama *dataset* yang akan digunakan pada sistem yang akan dibuat nantinya. Dataset akan dicari pada UCI Machine Learning Repository, dan KEEL *dataset repository.*

3. **Analisis Kebutuhan Sistem dan Perancangan sistem**

   Analisis terhadap kebutuhan sistem akan dilakukan pada tahap ini untuk nantinya akan digunakan untuk proses perancangan sistem untuk melakukan proses *feature selection* menggunakan metode ReliefF.

4. **Implementasi Rancangan Sistem**

   Pada tahap ini, akan dibangun sistem yang sudah dirancang untuk melakukan proses seleksi fitur menggunakan algoritma ReliefF.

5. **Pengujian sistem dan analisis hasil pengujian.**

   Tahapan ini merupakan implementasi hasil perancangan sistem yang telah dibuat. Dari hasil yang dikeluarkan oleh sistem akan dilakukan analisis.

6. **Tahap pembuatan laporan.**

   Pada tahap ini akan dilakukan penyusunan laporan sebagai dokumentasi apa yang selama ini telah dikerjakan dengan mengikuti aturan penulisan yang diberikan oleh institusi.

Bab 5 Penutup

5.1. Kesimpulan

1. Parameter yang mempengaruhi kinerja dari algoritma adalah pemilihan jumlah instan atau jumlah iterasi yang akan dilakukan, jumlah tetangga terdekat untuk mencari nilai *near hit* dan *near miss*.
2. Pemilihan parameter *m* optimum berada pada kisaran 10 – 50, sedangkan parameter *k* optimum pada 5 – 10.
3. Kualitas *dataset* hasil pemilihan fitur dengan menggunakan algoritma ReliefF sangat bergantung pada kualitas sampel acak yang terpilih oleh sistem.
4. Algoritma ReliefF mampu untuk menangani *dataset* dengan tipe data numerik (real), diskrit (nominal dan kategoris), serta gabungan antara numerik dan diskrit melalui proses pencarian tetangga terdekatnya menggunakan perhitungan pencarian tetangga terdekat pada tipe data numerik dan diskrit yang dijumlahkan.
5. Berdasarkan perhitungan nilai performansi pada *classifier* IB1 dan Naïve Bayes, secara garis besar ReliefF mampu memberikan peningkatan performansi hasil klasifikasi pada keduanya.

5.2. Saran

1. Bisa menggunakan *dataset* lain yang mengandung *missing value* atau *noise*.
2. Gunakan teknik lain dari keluarga Relief yaitu seperti RReliefF atau Iterative Relief.

Daftar Pustaka

[1] Kononenko, I. 1994. Estimating attributes: Analysis and Extensions of RELIEF. *European Conference on Machine Learning* (pp. 171- 182)

[2] Kira, K. and L. A. Rendell .1992. *A Practical Approach to Feature Selection*. 9th International Workshop on Machine Intelligence, Aberdeen, Scotland, Morgan-Kaufman.

[3] Robnik-Sikonja, M. Kononenko, I . 2003. *Theoretical and Empirical Analysis of ReliefF and RReliefF*. Machine Learning Journal 53:23-69

[4] Liu, Huan dkk. 2010. *Feature Selection : An Ever Evolving Frontier in Data Mining*. JMLR : Workshop and Conference Proceedings 10: 4-13.

[5] Draper, Bruce., Carol K, Jose Bins. 2003. *Iterative Relief*. Workshop on Learning in Computer Vision and Pattern Recognition, Madison, WI.

[6] Liu, Huan.Hiroshi Motoda. 1998. *Feature Selection for Knowledge Discovery and Data Mining*. Massachusetts : Kluwer Academic Publisher.

[7] Zafra, Amelia., Pechenizkiyz, M. & Ventura, S. 2010. *Feature Selection is the ReliefF for Multiple Instance Learning.* Intelligent Systems Design and Applications (ISDA) 2010 Conference.

[8] Dash, M., H Liu. 1997. *Feature Selection for Classification.* Intelligent Data Analysis pp. 131 – 156.

[9] Kononenko, Igor. Marko Robnik-Sikonja, Uros Pompe. 1996 . *ReliefF for estimation and discretization of attributes in classification, regression and ILP problems*. In A. Ramsay (ed.): Artificial Intelligence: Methodology, Systems, Applications: Proceddings of AIMSA'96, pp.31-40, IOS Press.

[10] Kononenko, Igor. 2005. *Evaluating the Quality of Attributes.*Advanced Course in Artificial Intelligence(ACAI)-05 summer school, IJS, Ljubljana. Pages 1-14.

[11] Liu, Y. and M Schumann. 2005. *Data mining feature selection for credit scoring models.*Journal of the Operational Research Society.

[12] Han, Jiawei. Micheline Kamber. 2006. *Data Mining Concepts and Techniques 2nd Edition*. Morgan Kaufmann.

[13] Liu, Huan. Hiroshi, Motoda. 2008. *Computation Methods of Feature Selection.* Chapman & Hall/CRC.

[14] Pyle, Dorian. 1999. *Data Preparation for Data Mining*. San Francisco:Morgan Kaufmann Publisher, Inc.

[15] Olson, David L. Delen, Dursun ”Advanced Data Mining Techniques” Springer; 1 edition (February 1, 2008), page 138

[16] W. Aha, David, Denis Kibler, Marc K.1991.*Machine Learning : Instance-Based Learning Algorithms*.Boston: Kluwer Academic Publishers.

[17] Arti, Prima., Imelda Atastina, Kemas Rahmat.2011.*Analisis Penanganan Missing Value dengan Metode Robust Least Squares Estimation with Principal Component (RLSP)*. Bandung:Fak Informatika IT Telkom.