ANALISIS DAN IMPLEMENTASI ALGORITMA RANDOM FOREST SEBAGAI SEBUAH CLASSIFIER DALAM DATA MINING

Rahmi Fitriani Ab[1], Moch. Arif Bijaksana[2], Adiwijaya[3]

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

---

Abstrak

Kata Kunci :

Abstract

Keywords :

---

1. Pendahuluan

1.1 Latar belakang

Seiring dengan perkembangan teknologi dalam hal pengumpulan dan penyimpanan data, maka munculah suatu kebutuhan untuk dapat menghasilkan informasi dari data yang telah ada tersebut. Dengan adanya ukuran data yang sangat besar diharapkan informasi penting yang didapat pun akan semakin banyak.

Salah satu cara untuk mendapatkan informasi atau pola dari kumpulan data yang besar adalah dengan menggunakan teknik-teknik dalam data mining. Ada banyak teknik dalam data mining untuk menghasilkan informasi atau pola dari sekumpulan data. Salah satu teknik tersebut adalah klasifikasi.

Random Forest merupakan salah satu algoritma klasifikasi dengan tingkat akurasi yang baik [1]. *Random Forest* merupakan sebuah metode *ensemble* yang terdiri dari beberapa pohon keputusan sebagai *classifier*. Kelas yang dihasilkan dari proses klasifikasi ini diambil dari kelas terbanyak yang dihasilkan oleh pohon-pohon keputusan yang ada pada *Random Forest.* Dengan melakukan voting pada pohon-pohon keputusan yang tersedia membuat akurasi dari *Random Forest* meningkat.

1.2 Perumusan masalah

Permasalahan yang dijadikan objek penelitian dalam Tugas Akhir ini adalah bagaimana mengimplementasikan algoritma klasifikasi *Random Forest*. Selain itu masalah lainnya adalah bagaimana melakukan analisa performansi dari algoritma klasifikasi *Random Forest*. Performansi yang dianalisa tersebut adalah akurasi dari prediksi yang dihasilkan.

Ruang lingkup dari Tugas Akhir ini adalah :

1. *Dataset* yang digunakan adalah *dataset* sintetik yang dihasilkan oleh data generator ataupun *dataset* yang berasal dari *UCI Machine Learning Repository*. *Dataset* yang dipakai berjumlah delapan *dataset*.
2. *Dataset* yang digunakan adalah *dataset* yang seluruh atributnya bertipe numerik.
3. Tidak menangani tahap *pre-processing*.
4. Pengujian dilakukan dengan parameter akurasi.
5. Inputan yang diterima berupa tabel yang telah tersedia di dalam *database*.

1.3 Tujuan

Tujuan Tugas Akhir ini adalah :

1. Merancang dan membangun perangkat lunak untuk klasifikasi data dengan algoritma *Random Forest.*
2. Menganalisa performansi dari algoritma *Random Forest.* Performansi yang dianalisa tersebut adalah akurasi dari prediksi yang dihasilkan serta membandingkannya dengan beberapa algoritma klasifikasi lain, yakni Naive Bayes, AdaboostM1, RepTree dan Bagging.

1.4 Metodologi penyelesaian masalah

Metodologi yang digunakan dalam Tugas Akhir ini antara lain :

1. Studi Literatur, yang dilakukan dengan mempelajari beberapa literatur (makalah, buku ataupun jurnal) yang berkaitan dengan *data mining*, klasifikasi terutama yang berkaitan dengan *random forest* dan metode untuk menganalisa performansi algoritma klasifikasi tersebut.
2. Pengumpulan dan analisis data yang digunakan untuk mendukung implementasi dan analisis algoritma *random forest.*
3. Analisis kebutuhan dan perancangan perangkat lunak yang akan dibangun. Dalam hal ini digunakan metode berorientasi objek.
4. Implementasi dari rancangan perangkat lunak yang telah dibuat, sehingga dihasilkan perangkat lunak yang dapat digunakan untuk mengklasifikasikan data dengan menerapkan algoritma *random forest*. Dalam pembuatan perangkat lunak ini digunakan bahasa pemograman Delphi 7, dan Microsoft Office Access.
5. Pengujian dan analisa hasil, melakukan uji kebenaran klasifikasi dan mengukur hasil implementasi berdasarkan parameter akurasi yang dihasilkan serta membandingkannya dengan beberapa algoritma klasifikasi lain yakni Naive Bayes, AdaboostM1, RepTree dan Bagging. Untuk membandingkan dengan algoritma klasifikasi lain tersebut, digunakan perangkat lunak WEKA.
6. Pengambilan kesimpulan dan penyusunan laporan Tugas Akhir.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Kesimpulan yang dapat diambil dari penelitian Tugas Akhir ini adalah :

1. *Random Forest* merupakan salah satu algoritma yang cukup efektif dalam melakukan klasifikasi. Ini terlihat dari grafik persentase error yang membandingkan dengan algoritma klasifikasi lain. *Random Forest* memperlihatkan persentase error yang bersaing dengan algoritma lain, dengan di beberapa data memiliki persentase error terkecil.
2. Pada *Random Forest*, ada suatu batasan jumlah pohon minimal yang harus dibangun, sehingga pada jumlah tersebut seluruh data sudah diklasifikasikan. Dimana jumlah tersebut sangat tergantung dengan masing-masing data. Jumlah atribut pemecah mempengaruhi jumlah pohon minimal untuk tiap data.
3. Jumlah pohon memberikan pengaruh yang besar terhadap tingkat akurasi. Dimulai pada jumlah pohon minimal, peningkatan jumlah pohon, meningkatkan akurasi yang dihasilkan. Ada suatu batas akurasi terbaik, dimana setelah akurasi itu dicapai, meskipun jumlah pohon ditambah akurasi akan tetap stabil.
4. Akurasi yang dihasilkan dipengaruhi oleh jumlah atribut pemecah. *Random Forest* dengan menggunakan jumlah atribut pemecah sama dengan jumlah atribut yang ada akan memberikan akurasi yang rendah.
5. Untuk tiap data, jumlah atribut pemecah yang menghasilkan akurasi terbaik berbeda-beda. Nilai atribut pemecah ini berada di sekitar nilai perkiraan atribut pemecah yang dihitung menggunakan persamaan 2.1.
6. Perbedaan jumlah pohon dalam pembangunan *Random Forest* tidak mengubah nilai OOB *error rate* untuk setiap pohon dalam *random forest.* Sedangkan perbedaan jumlah atribut pemecah, akan mengubah nilai OOB *error rate* untuk tiap pohon yang ada. Hal ini terkait dengan fungsi acak yang dibangkitkan, sesuai dengan jumlah atribut pemecah.
7. *Random Forest* memiliki kelemahan dalam hal kestabilan akurasi yang dihasilkan. Dengan parameter input dan data yang sama, untuk eksekusi lebih dari satu kali yang dilakukan berurutan akan menghasilkan akurasi yang berbeda-beda. Hal ini terkait dengan fungsi acak yang dibangkitkan untuk melakukan pemilihan baris data dan pemilihan kandidat atribut pemecah secara acak.

5.2 Saran

Saran untuk pengembangan Tugas Akhir ini adalah :

1. *Random Forest* dapat dikembangkan sehingga bisa menangani data *imbalance class* dengan lebih memfokuskan pada perbaikan prediksi untuk data dengan kelas minor. Sehingga untuk pembentukan pohon, data yang digunakan terutama data-data dengan kelas minor.
2. *Random Forest* dapat dikembangkan sehingga bisa menangani berbagai tipe data yang diinputkan, termasuk menangani data yang memiliki *missing value.*
3. *Random Forest* dapat dikembangkan sehingga bisa melakukan *outlier detection.*

Daftar Pustaka

[1]	Breiman, Leo and Cutler, Adele. Random Forest. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm, didownload pada tanggal 13 November 2006
[2]	Breiman, Leo, 2001. *Random Forests*. University Of California At Berkeley
[3]	Breiman, Leo. 2003. RF / tools : *A Class of Two-eyed Algorithms*. University Of California At Berkeley.
[4]	Breiman, Leo, 2004. *Consistency For A Simple Model Of Random Forests*. University Of California At Berkeley.
[5]	Chawla V Nitesh, "*C 4.5 and Imbalanced Data sets : Investigating the effect of sampling method , probabilistic estimate, and decision tree structure"*, Canada, 2003.
[6]	Dharmastuti, Bhakti, 2007. Analisis Penggunaan C&RT sebagai *Binary Recursive Partitioning* pada *Decision Tree.* STT Telkom, Bandung.
[7]	Gehrke, Johannes. *Advances in Decision Tree Construction*. Cornell University, Visconsin, Madison.
[8]	Han, Jiawei and M. Kamber, 2001. *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann, 2001.
[9]	Kurniati, Angelina Prima, 2005. Analisis dan implementasi algoritma public sebagai sebuah *classifier* pohon keputusan yang *scalable* dalam data mining. STT Telkom, Bandung.
[10]	Liaw, Andy and M. Wiener, 2002. *Classification and Regression by Random Forest*
[11]	Pang-Ning Tan, M. Steinbach and V. Kumar 2006. *Intoduction to Data Mining*
[12]	Stephens, Rod. http://www.vb-helper.com/dart1.htm#Quicksort, didownload pada tanggal 15 Juni 2007
[13]	Suhendar, A. dan Gunadi, Hariman, "Visual Modelling Menggunakan UML dan Rational Rose", Informatika Bandung, 2002.
[14]	http://www.ics.uci.edu/~mlearn/MLRepository.html