ANALISA DAN IMPLEMENTASI METODE ENSEMBLE RANDOM FOREST DAN ALGORITMA C4.5 UNTUK KLASIFIKASI DATA EMAIL SPAM

Ridha Ramadhansyah[1], Arie Ardiyanti Suryani[2], Kemas Rahmat Saleh Wiharja[3]

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

Abstrak

Berkembangnya kemajuan teknologi komunikasi saat ini memudahkan orang-orang untuk berkomunikasi. Salah satu teknologi komunikasi yang sering digunakan adalah email. Email sendiri adalah suatu metode pertukaran pesan pesan digital antara satu orang pembuat pesan ke satu atau banyak penerima. Email sendiri sekarang sering disalah gunakan untuk kepentingan pribadi, kelompok atau organisasi. Terkadang seseorang pemilik email menerima pesan yang tidak diinginkan dari orang yang tidak dikenal. Email ini disebut sebagai email spam. Salah satu cara untuk menangani ini adalah melakukan pemfilteran email spam berdasarkan konten. Metode learning pohon keputusan dengan algortima C4.5 dapat digunakan untuk menangani permasalahan tersebut.Untuk mendapatkan kinerja pohon keputusan yang baik dapat menggunakan metode ensemble Random Forest. Random Forestmengunakan kombinasi dari beberapa pohon keputusan dengan inputan jumlah feature yang akan digunakan pada pohon keputusan. Penggunaan jumlah feature yang tepat akan menghasilkan nilai akurasi yang baik dalam memfilter data spam atau ham.

Kata Kunci : pemfilteran email, pohon keputusan, algoritma C4.5, metode ensemble, random forest, akurasi.

Abstract

The development of communication technology nowadays make people can communicate easily. One of the communication technology that is often used is email. Email is a method of digital messages exchange between person who make the message to one or many recipients. Email nowadays is often misused for personal interest, groups or organizations. Sometimes the owner of an email received unwanted messages from the strangers . This kind of email is called as spam email. One way to handle this problem is doing spam filtration based on the content. C4.5 algorithm decision tree learning method can be used to handle these problems. To get a good performance of decision tree, Random Forest ensemble method can be used. Random Forest is using a combination of several decision trees to input the number of features that will be used in the decision tree. The using of the appropriate amount of features will produce good accuracy in filtering spam data or ham.

Keywords : email filtering,decision tree, C4.5 algorithm, ensemble method, random forest, accuracy.

BAB I
PENDAHULUAN

1.1 Latar BelakangMasalah

Spam atau bisa juga berbentuk *junk mail* adalah penyalahgunaan sistem pesan elektronik (termasuk media penyiaran dan sistem pengiriman digital) untuk mengirim berita iklan dan keperluan lainnya secara massal.Setiap spam yang diterima memakan waktu dan tenaga si penerimanya untuk membaca, menyortir, menghapus, berusaha menolak di kemudian hari. Spam juga bisa memenuhi mailbox, mengakibatkan mailserver sibuk, dan memperlambat layanan lainnya.Walaupun email spam dapat di tangani dari para penyedia layanan internet dari sisi email header terkadang masih ada spam yang lolos dari pengawasan dari sisi konten email.

Untuk melakukan klasifikasi terhadap email spam dapat dilakukan dengan pohon keputusan. Pada tugas akhir ini, akan dibangun sistem yang menerapkan algoritma klasifikasi pohon keputusan C4.5 untuk mengklasifikasikan email spam dan ham(bukan spam). Kemampuan algoritma C4.5 untuk mem-breakdown proses pengambilan keputusan yang kompleks menjadi lebih simple menghasilkan pengambilan keputusan yang lebih menginterpretasikan solusi dari permasalahan.

Decision treemerupakan algoritma learning yang *unstable*, perubahan kecil terhadap training set mengakibatkan perubahan yang besar pada learned *classifier*[1]. Salah satu cara untuk mengatasinya dengan metode ensemble, membentuk beragam *classifier* dengan memanipulasi data trining[4]. Untuk itu pada tugas akhir ini algoritma C4.5 menjadi *classifier* dalam metode ensemble.Pendekatan ensemble yang digunakan adalah random forest.

Tugas akhir ini menganalisaperformasi random forestdengan algoritma klasifikasi algoritma C4.5 dalam kasus pengklasifikasian terhadap konten dari spam email.

1.2Perumusan Masalah

Dari latar belakang di atas maka masalah-masalah yang dihadapi, yaitu :

1. Bagaimana mengimplementasikan email spam filtering berdasarkan konten dengan mengunakan metode decision tree.
2. Bagaimana cara memanipulasi data training sebelum digunakan pada algoritma klasifikasi.
3. Bagaimana cara mendapatkan model yang terbaik dari metode Random Forest dengan classifier dari Algoritma C4.5.

Batasan dari permasalahan adalah :

1. Data yang digunakan telah melalui tahap*preprocessing*.
2. Data yang diterima sistem dalam bentuk tipe *continuous*.
3. Data inputan berasal dari email dalam bahasa Inggris yang telah di-*preprocessing*. Karena dari risetsebagian besar spam email yang menyebar mengunakan bahasa Inggris[6].

1.3Tujuan

Tujuan yang ingin dicapai dalam pembuatan tugas akhir ini adalah sebagai berikut :

1. Mengimplementasi metode decision tree menggunakan algoritma C4.5 pada permasalahan konten email spam.
2. Menerapakan metode ensemble yaitu random forest untuk memaipulasi data yang akan di gunakan pada *classifier*.
3. Menganalisamodel yang menerapkan metode random forest dan C4.5 sebagai *classifier*dan menguji mengunakan data testing.

1.4 Hipotesis

Pengklasifikasian email spam menggunakan random forest sebagai metode ensemble dengan *classifier*-nya dari algoritma C4.5 menghasilkan klasifikasi email spam yang memiliki akurasi yang baik (minimal 60%).

1.5Metodologi Penyelesaian Masalah

Metode yang digunakan untuk menyelesaikan masalah yaitu :

1. Studi Pustaka
   Bahan bahan dalam penyelesaian masalah didapat dari literaturbaik itu jurnal, buku-bukuatau informasi lain yang relevan yang berhubungan dengan Algoritma C4.5, metode ensemble, random forest,text categorization, email filtering, dan decision tree.

2. Pengumpulan data
   Data digunakan didapat dari UCI repository dengan dataset bernama Spambase dengan jumlah instan 4601 dan atribut 57.

3. Implementasi sistem
   Pembentukan sistem yang dapat membentuk model yang menerapkan metode Random Forest dengan *classifier* dari algoritma C4.5.

4. Pengujian
   Model yang terbentuk diujikan mengunakan data uji yang telah dipersiapkan untuk melihat kesesuaian hasil prediksi yang dibuat oleh model.

5. Analisis
   Analisis dilakukan tehadap hasil pengujian dari sistem sehingga dapat dibentuk suatu kesimpulan.

6. Perumusan kesimpulan penyelesaian masalah
   Merumuskan kesimpulan tehadap hasil pengujian pada sistem untuk menyelesaikan masalah.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari hasil analisis diatas dapat ditarik kesimpulan sebagai berikut:

1. Waktu pembentukan model semakin lama saat jumlah feature yang digunakan pada model lebih banyak.
2. Dari analisa mengunakan data email spam dari UCI repository, untuk mendapatkan model optimal yang menerapkan metode Random Forest dengan algoritma klasifikasi C4.5 adalah model dengan inputan jumlah feature 4.
3. Nilai estimasi OOB error rate merepresentasikan nilai akurasi model dan dapat digunakan untuk melihat kualitas model dari input jumlah feature.

5.2 Saran

1. Penggunaan metode ensemble yang lain seperti bagging atau boosting untuk memanipulasi data training untuk algoritma C4.5.

DAFTAR PUSTAKA

[1] Breiman, Leo. 1996. Heuristics Of Instability And StabilizationIn Model Selection.University of California.USA

[2] Breiman, Leo. 2001. Random Forest. Statistics Department, University of California.

[3] Breiman,Leo.2004.Random Forest . http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm. Diakses pada tanggal 27 Mei 2010.

[4] Dietterich, Thomas G. 2000. Ensemble Methods in Machine Learning. Oregon State University. USA.

[5] Han, jiawei. 2006. Data Mining: Concepts and Techniques. Morgan Kaufmann, University of Illinois. USA.

[6] MarissaVicario. 2009.The Language of Spam: Spammers do their Homework before Spamming Specific Regions. http://www.symantec.com/connect/blogs/language-spam-spammers-do-their-homework-spamming-specific-regions-0. Diakses pada tanggal 21 November 2010.

[7] Quilan, Ros, Ron Kohavi. 1999. Decision Tree Discovery. http://ai.stanford.edu/~ronnyk/treesHB.pdf .Diakses pada tanggal 3Februari 2011.

[8] Sebastian, Fabrizio. 2005. Text Categorization. University Padova. Italy.

[9] Tan, Pang-Ning. 2006. Introduction to Data Mining. Chapter 4. Classification: Basic Concepts, Decision Trees, and Model Evaluation. Addison-Wesley Logman Publishing. USA.

[10] Techsoup .2006. Ten Spam-Filtering Methods Explained. http://www.techsoup.org/learningcenter/internet/page6028.cfm . Diakses pada tanggal 28 Mei 2011.

[11] Saleh, Rachmad.2008. Spam dan Hijacking Email. Andi Publisher.Jakarta.

[12] Dmoz. 2011. Email Filtering. http://www.dmoz.org/Computers/Internet/E-mail/Spam/Filtering/. Diakses pada tanggal 2 Juli 2011.

[13] Commtouch. 2010.Internet Threats Trend Report Q1 2010. USA.