IMPLEMENTASI DAN ANALISA E-MAIL SPAM FILTERING MENGGUNAKAN GRANULAR SUPPORT VECTOR MACHINES – RECURSIVE FEATURE ELIMINATION (GSVM-RFE)

Ririn Zulandra[1], Tri Brotoharsono[2], Erda Guslinar Perdana[3]

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

Abstrak

E-mail spam merupakan pesan yang tidak diminta yang dikirim ke sejumlah penerima melalui e-mail dalam jumlah yang besar. Contoh kiriman e-mail yang merupakan spam seperti iklan, tawaran untuk bergabung ke MLM, undian, informasi palsu, phishing, dan penipuan. Dengan adanya e-mail spam ini mengakibatkan pengguna e-mail membutuhkan waktu yang lebih banyak untuk membaca dan memutuskan apakah e-mail yang diterima tersebut adalah spam atau bukan.

Dalam tugas akhir ini, dibangun sebuah sistem e-mail spam filtering dengan menggunakan algoritma Granular Support Vector Machine-Recursive Feature Elimination (GSVM-RFE). GSVM-RFE akan menghapus kata-kata yang irrelevant, redundant, noisy yang terdapat pada e-mail dalam granules yang berbeda dan memilih kata-kata yang informatif untuk digunakan dalam klasifikasi. Dengan cara kerja yang demikian GSVM-RFE dapat menghasilkan akurasi sebesar 98% dengan 8192 kata yang dihasilkan

Kata Kunci : e-mail spam filtering, granular computing, fuzzy C-means clustering, recursive feature elimination, relevance index, support vector machines

Abstract

E-mail spam are unsolicited messages that sent to multiple recipients via e-mail in large numbers. The example of e-mail spam is advertisement e-mail, bid to join the MLM, sweepstakes, fake information, phishing and fraud. With this e-mail spam, an e-mail users need more time to read and decide whether the e-mail received is spam or not.

In this work, we built an e-mail spam filtering system using Granular Support Vector Machine-Recursive Feature Elimination (GSVM-RFE) algoritm. The GSVM-RFE separately eliminates irrelevant, redundant, or noisy word at email in different granules at different stages and selects highly informative word that used in classification. With that kind of word GSVM-RFE can produce 98% accuracy with 8192 words

Keywords : Keywords: e-mail spam filtering, granular computing, fuzzy C-means clustering, recursive feature elimination, relevance index, support vector machines,

1. Pendahuluan

1.1 Latar belakang

E-mail spam merupakan pesan yang tidak diminta yang dikirim ke sejumlah penerima melalui *e-mail* dalam jumlah yang besar. Contoh kiriman *e-mail* yang merupakan *spam* seperti iklan, advertensi, tawaran untuk bergabung ke MLM, undian, informasi palsu, *phishing*, dan penipuan [8]. Dengan adanya *e-mail spam* ini mengakibatkan pengguna *e-mail* membutuhkan waktu yang lebih banyak untuk membaca dan memutuskan apakah *e-mail* yang diterima tersebut adalah *spam* atau bukan. Seperti pada laporan hasil studi lembaga riset ICF International yang menyatakan bahwa pada tahun 2008, 80% energi dihabiskan oleh pengguna untuk menghapus *spam* dan mencari *e-mail* bukan *spam* [9]. Untuk menghadapi masalah ini banyak pendekatan yang diterapkan diantaranya *extensions of e-mail protocols, certification of e-mail server, e-mail spam filtering,* dan *legislation*, diantara pendekatan-pendekatan tersebut, *e-mail spam filtering* adalah solusi yang lebih realistis dalam implementasinya [12].

E-mail spam filtering adalah sebuah mekanisme yang digunakan untuk memisahkan *e-mail spam* dengan *e-mail* bukan *spam* secara otomatis [11]. Salah satu teknik yang digunakan dalam *E-mail spam filtering* yaitu teknik klasifikasi [5]. Klasifikasi adalah salah satu teknik dalam data *mining* yang digunakan untuk memprediksi kelompok keanggotaan(*class*) dari setiap *instance* data. Salah satu teknik *machine learning* untuk klasifikasi yang paling sering digunakan untuk membangun *e-mail spam filtering* adalah SVM.

Support Vector Machines (SVM) adalah salah satu teknik *machine learning* yang digunakan untuk klasifikasi data dan umumnya diimplementasikan untuk menangani *dataset* yang hanya memiliki dua kelas [17][15][2]. SVM banyak digunakan dalam pembangunan sistem *e-mail spam filtering* karena telah terbukti efisien dan akurasinya yang tinggi [20][13]. Akan tetapi performansi dari SVM akan menurun ketika *dataset* yang digunakan mengandung banyak atribut sehingga ketika dipetakan ke dalam ruang vektor menimbulkan *curse of dimensionality*. Kondisi lain yang mengkibatkan performansi SVM menurun yaitu jika suatu dataset jumlah *class* yang satu dengan yang lainnya sangat berbeda jauh. Kondisi tersebut menyebabkan *class* yang sedikit tersebut dianggap *outlier*. Untuk mengatasi hal tersebut, maka banyak dilakukan modifikasi terhadap SVM yang bertujuan untuk meningkatkan efektifitas dan efisiensi SVM[5].

Salah satu modifikasi pada SVM yaitu penggunaan paradigma *granular computing* dan teori statistik, yang kemudian penggabungannya disebut *Granular Support Vector Machines* (GSVM) [18]. Dalam pembangunan GSVM sendiri, banyak algoritma-algoritma yang diterapkan sesuai dengan tujuan pengimplementasian SVM itu sendiri. Salah satunya adalah algoritma *Recursive Feature Elimination* (RFE). Tahap awal algoritma ini akan menghilangkan *feature-feature* yang *irrelevant* dan *redundant* secara iteratif sehingga mencapai jumlah tertentu. Dan dari hasil tersebut akan di lakukan pemilihan *feature* yang paling merepresentasikan *email spam*. [18].

Pada tugas akhir ini akan dibangun sebuah sistem *e-mail spam filtering* menggunakan *Granular Support Vector Machine* dengan *Recursive Feature*

Elimination (RFE) karena algoritma ini dapat menghasilkan *feature-feature* yang lebih akurat atau informative untuk pengklasifikasian *email spam* [18].

1.2 Perumusan masalah

Berdasarkan latar belakang tersebut, maka dapat dirumuskan permasalahan sebagai berikut:

1. Bagaimana membangun sebuah sistem *e-mail spam filtering* dengan *Granular Support Vector Machines – Recursive Feature Elimination (RFE).*
2. Bagaimana akurasi dan efisiensi waktu *Support Vector Machines* dimodifikasi dengan *granular computing* dengan algoritma *Recursive Feature Elimination* terhadap *e-mail spam filtering* yang akan dibangun.

Terdapat beberapa batasan masalah dalam penelitian tugas akhir ini, antara lain :

1. Data yang digunakan adalah data yang tidak mengandung gambar dan bahasa yang digunakan adalah bahasa inggris.
2. Sistem yang dibangun adalah aplikasi yang berdiri sendiri (*stand alone application*) dan tidak diimplementasikan dalam *e-mail server*.

1.3 Tujuan

Tujuan dari penelitian tugas akhir ini adalah:

1. Mengimplementasikan teknik *Granular Support Vector Machine with Recursive Feature Elimination (RFE)* untuk mengklasifikasikan *email* berdasarkan kriteria *spam* atau bukan *spam*.
2. Melakukan analisa akurasi pada sistem yang telah dibangun dengan parameter *sensitivity*, *specitifity*.

1.4 Metodologi penyelesaian masalah

Metode yang digunakan untuk menyelesaikan tugas akhir ini adalah :

1. Identifikasi masalah, yakni dengan melakukan identifikasi terhadap permasalahan yang ada.
2. Studi Literatur, yakni mempelajari referensi dan literatur, baik berupa makalah, jurnal, maupun buku yang relevan yang membahas tentang *Granular Support Vector Machine with Recursive Feature Elimination (RFE)*.
3. Mempersiapkan data set yang akan digunakan untuk testing dan training dengan melakukan preprocessing terhadap data set tersebut.
4. Pembuatan desain sistem *e-mail spam filtering* dengan *Granular Support Vector Machine with Recursive Feature Elimination (RFE).*
5. Implementasi (*Coding*), yaitu mengimplementasikan perancangan menjadi sistem *email spam filtering* dengan menerapkan *Granular Support Vector Machine with Recursive Feature Elimination (RFE).*
6. Training dan testing sistem, melakukan pelatihan dan pengujian pada sistem dengan menggunakan data training dan data testing.
7. Analisa hasil, melakukan analisa hasil dari sistem dengan cara membandingkan hasil klasifikasi data testing dengan data jawaban sebenarnya.

8. Pembuatan laporan, mendokumentasikan semua tahap metodologi penyelesaian masalah menjadi suatu laporan yang nantinya dapat dikembangkan sesuai perkembangan jaman dan dapat dimanfaatkan sebagaimana mestinya.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil pengujian dan analisis yang telah dilakukan, dapat diambil beberapa kesimpulan sebagai berikut:

1. Metode GSVM-RFE dapat mengklasifikasikan *dataset* yang digunakan dengan akurasi yang tinggi, yaitu 99,6322% untuk data *training* dan 98% untuk data testing, dengan nilai parameter positif *filtering threshold* = 0.9,negative *filtering threshold* = 0.5, jumlah *granul* = 2, *filter factor* = 0.1 dan C = 0,9.
2. Nilai *sensitivity* dan *specificity* yang dihasilkan oleh metode GSVM-RFE dengan parameter terbaik adalah seimbang yaitu 0.966 dan 0.994. Ini bearti metode GSVM-RFE dapat mengklasifikasikan *e-mail* berdasarkan *spam* atau tidak dengan seimbang.
3. Semakin besar nilai parameter C yang digunakan, semakin besar juga pinalty eror sehingga eror yang dihasilkan kecil dan akurasi tinggi. Namun nilai C yang terlalu besar dapat mengakibatkan akurasi menurun karena hyperplane terlalu ketat sehingga data salah terklasifikasikan. Pada dataset *e-mail* yang digunakan nilai C yang diambil adalah 0.9.
4. Metode GSVM-RFE yang digunakan sudah cukup bisa menebak dengan benar data *spam* yang diambil dari *account* GMail penulis , dengan 7 *e-mail* dari 10 *e-mail* yang dengan benar di klasifikasikan. Walaupun masih terdapat kesalahan. Kesalahan ini dikarenakan data *training* yang digunakan kurang mewakili data *testing* yang diambil dari GMail. Selain itu, sistem yang digunakan oleh GMail berbeda dengan yang digunakan oleh penulis.

5.2 Saran

Beberapa saran yang diberikan antara lain:

1. Penggunaan kernel lain pada SVM yang digunakan untuk klasifikasi data mungkin akan memberikan hasil yang berbeda. Namun perlu diperhatikan apakah jumlah atribut yang digunakan akan menimbulkan *curse of dimensionality* atau tidak.
2. Penggunaan data yang lebih mewakili data yang terdapat di GMail atau *Yahoo!* sebagai data *training* mungkin akan membuat sistem yang dibangun menjadi lebih sesuai dengan data *spam* yang beredar di dunia maya.
3. Penggunaan metode GSVM-RFE dimana SVM-RFE yang digunakan diganti dengan SVM-RFE *two stage*.
4. Penggunaan metode GSVM-RFE untuk menangani berbagai macam data *spam* lainnya, misalnya data *spam* yang mengandung gambar didalamnya.

Referensi

[1] B.Enrico, B.Anton, 2007,*A Survey of Learning-Based Techniques of Email Spam Filtering Enrico Blanzieri, Anton Bryl*, University of Trento, Italy

[2] B. Steve, 2003, *Support Vector Machines*, Department of Computer Science and Artificial Intelligence, University of Malta. In Proceedings of the First Computer Science Annual Workshop (CSAW) 2003

[3] "Fuzzy C-Means Clustering" http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html diakses pada tanggal 20 Oktober 2010

[4] Google, "Gmail uses Google's innovative technology to keep spam out of your inbox", http://www.google.com/mail/help/fightspam/spamexplained.html diakses pada tanggal 29 november 2011

[5] G. Sudipto, M. Adam, M. Nina, M. Rajeev, O. Liadan, 2003, *Clustering Data Streams: Theory and Practice*, Radical Eye Software, pages 1-4

[6] H. Eyke, 2008, Granular Computation in Machine Learning and Data Mining. In P. Witold, S. Andrzej, K Vladik, (editors), Handbook of Granular Computing, John Wiley & Sons, pages 889-907

[7] H.Jiawei,K.Micheline,2006,*Data Mining Concepts and Techniques*,Morgan Kaufmann Publishers

[8] http://techscape.co.id/hosting/spam.ts diakses pada tanggal 19 Oktober 2010

[9] http://blog.indocisc.co.id/2009/07/15/kerugian-dikaitkan-dengan-lingkungan-akibat-spam/ diakses pada tanggal 19 Oktober 2010

[10] http://www.eweek.com/c/a/Security/Google-Adds-DKIM-for-Google-Apps-to-Address-Spam-243360/ diakses pada tanggal 29 november

[11] http://www.marketingterms.com/dictionary/EmailSpam-definition,information, sites, articles.htm diakses pada tanggal 19 Oktober 2010

[12] L. Shugang, C. Kebin, 2009, *Applications of Support Vector Machine Based on Boolean Kernel to Spam Filtering, Journal of Modern Applied Science Volume 3 Number 10*, School of Computer Science and Technology, North China Electric Power University

[13] M. Eirinaios, A. Ion, P. Georgios, S. George, S. Panagiotis, 2002, Filtron: *A Learning-Based Anti-Spam Filter, In Proceedings of The first Conference on Email and Anti-Spam (CEAS)*, Mountain View, California

[14] Nugroho, A.S., Witarto, A.B., Handoko, D., "*Application of Support Vector Machine in Bioinformatics*", Proceeding of Indonesian Scientific Meeting in Central
Japan, December 20, 2003, Gifu-JapanAnto Satriyo Nugroho, Arief Budi Witarto, Dwi Handoko

[15] R. G. Steve, 1998, *Support Vector Machines for Classification and Regression*, Technical Report Faculty of Engineering, Science and Mathematics, University of Southampton

[16] S.G.Budhi, G.Ibnu, Y.Ferry, *Algoritma Porter Stemmer for Bahasa Indonesi Untuk Pre-Processing Text Mining Berbasis Metode Market Basket Analisis*, Dosen UK Petra Jurusan Teknik Informatika

[17] S. N. William, 2006, *What is A Support Vector Machines?*, Nature Publishing Group, Nature Biotechnology Volume 24 Number 12, pages 1565-1567

[18] Tang Yuchun. (2006). *Granular Support Vector Machines Based on Granular Computing, Soft Computing and Statistical Learning*

[19] T. Yuchun, Z. Yan-Qing, C. Nitesh V., K. Sven, 2002, "SVMs Modeling for Highly Imbalanced Classification". Journal of latex class files

[20] W. Qiang, 2006, G. Yi, W. Xisolong, *SVM-Based Spam Filter with Active and Online Learning*, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001