KLASIFIKASI DOKUMEN WEB DENGAN MENGGUNAKAN SUPPORT VECTOR MACHINE (SVM)

Hiskia Edy Pasaribu[1], Imelda Atastina[2], Shaufiah[3]

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

Abstrak

Seiring dengan pertumbuhan situs di internet yang sangat pesat, perlu dilakukan penyusunan dan pengorganisasian dokumen web (webpage) agar memudahkan pencarian, pengelolaan, dan mendapatkan informasi sesuai dengan kebutuhannya. Proses klasifikasi adalah solusinya yaitu dengan menggunakan machine learning.

Dari pengujian sebelumnya, metoda machine learning yang digunakan adalah Support Vector Machine (SVM), tetapi yang menjadi salah satu kelemahannya adalah klasifikasi yang dihasilkan oleh classifier SVM itu tidak dapat diketahui apakah merupakan suatu dugaan atau jawaban yang pasti. Dalam tugas akhir ini akan dilakukan suatu pendekatan baru dalam mengklasifikasikan dokumen web menggunakan SVM agar klasifikasi yang dihasilkan menjadi reliable, yaitu dengan menerapkan Version Space (VS). Version space adalah sebuah pendekatan untuk mendapatkan klasifikasi yang reliable. Ide utamanya adalah membangun version space yang mengandung sekumpulan fungsi hipotesis. Rule dari version space yang disebut dengan unanimous voting rule, akan digunakan untuk menjamin bahwa jika suatu data baru diklasifikasikan maka data tersebut benar diklasifikasikan.

Dalam tugas akhir ini akan dilakukan analisis terhadap hasil training dan testing. Pengujian juga akan menerapkan feature selection. Hasil dari penelitian menunjukkan bahwa performansi paling optimal diperoleh saat menerapkan feature selection yaitu 82.35 % pada domain data pertama dengan hanya menggunakan 60-90 atribut dari total 175 atribut, dan 80 % pada domain data kedua dengan hanya menggunakan 30 atribut dari total 214 atribut.

Kata Kunci : klasifikasi, preprocesing, reliable, SVM, VS, webpage.

Abstract

Along with the growth of sites on the Internet that is very fast, needs to be done preparation and organization of web documents (webpages) for easier search, manage, and obtain information in accordance with their needs. The process of classification is the solution that is by using machine learning.

The method of machine learning that used previously is Support Vector Machine (SVM), but the one of weakness of SVM is the classification that produced by classifier of SVM can't be known whether an assumpsion or definite answer. In this final task will be done a new approach in classfying web documents using SVM in order to the classification that produced by classifier SVM become reliable, that is by applying the Version Space (VS). Version space is an approach to obtain a reliable classification. The key idea is to construct version space containing a set of hypotheses. Rule of the version space that called unanimous vote rule will be used to quarantee that if a new data classified then it correctly classified.

In this final assignment will be done an analysis to the results of training and testing. This task will also apply feature selection. Results of the task showed that the optimal performance obtained when applying feature selection that is 82.35% in the first domain by using only 60-90 attributes of total 175 attributes, and 80% in the second domain by using only 30 attributes of total 214 attributes.

Keywords : classification, preprocesing, reliable, SVM, VS, webpage.

1. Pendahuluan

1.1 Latar Belakang

Saat ini perkembangan data dalam dunia maya (*World Wide Web*) sangat pesat. Terdapat berjuta-juta halaman web yang dapat dilihat. Data yang bisa diamati dari berjuta-juta situs di internet juga sangat beragam, mulai dari data teks dalam berbagai format seperti html, pdf, dokumen word, text, dan sebagainya sampai kepada data gambar dan data audio/video dalam berbagai format. Bahkan saat ini file dengan ukuran yang besar dapat dengan mudah diambil hanya melalui koneksi internet. Dengan demikian, pada saat ini kebutuhan informasi dalam dunia internet semakin meningkat.

Seiring dengan pertumbuhan situs, kebutuhan yang banyak diperlukan para pengguna internet adalah bagaimana mendapatkan informasi yang spesifik dan sesuai dengan kebutuhan, mengingat bahwa ketika ukuran web tersebut bukan hanya semakin besar, tetapi isi web itu juga cepat berubah-ubah, bersifat heterogen, terdiri dari berbagai bahasa, dan lain sebagainya. Seiring dengan pertumbuhan situs di Internet yang sangat besar itu, perlu dilakukan penyusunan atau pengorganisasian dokumen web agar memudahkan para pengguna untuk melakukan pencarian, pengolahan dan mendapatkan informasi sesuai dengan kebutuhannya. Pengklasifikasian adalah salah satu bentuknya.

Klasifikasi yang dimaksud disini adalah suatu bagian teknik *Web Mining* yaitu *Web Content Mining* yang berfokus pada analisa *content* informasi teks yang tersimpan pada dokumen web. Teknik klasifikasi dokumen web ini kurang memberikan solusi jika jumlah dokumen webnya sangat banyak apalagi jika dilakukan secara manual. Maka klasifikasi dokumen web ini akan dilakukan dengan menggunakan *machine learning*.

Dari pengujian sebelumnya, metoda machine learning yang dipakai adalah menggunakan Support Vector Machine (SVM). Salah satu kelemahan dari SVM adalah tidak ada yang tahu apakah hasil klasifikasi yang dihasilkan oleh *classifier* SVM itu merupakan suatu dugaan atau suatu jawaban yang pasti, sebab classifier yang dihasilkan SVM, belajar dari pengalaman dan ekstraksi pengetahuan yang ada dalam database bertujuan untuk bisa mengklasifikasikan data baru, tetapi tidak bisa membedakan hasil jawaban apakah merupakan suatu dugaan atau suatu jawaban yang pasti, dengan kata lain hasil klasifikasinya tidak reliable. Dalam tugas akhir ini akan diterapkan suatu pendekatan baru dalam mengklasifikasian dokumen web dengan SVM untuk mendapatkan hasil klasifikasi yang reliable. Salah satu pendekatan yang baik untuk mendapatkan klasifiaksi yang reliable adalah dengan menerapkan *version space (VS)*. Dalam tugas akhir ini, SVM akan dikombinasikan dengan version space untuk mendapatkan hasil klasifikasi SVM yang reliable.

Version space adalah suatu pendekatan klasifikasi untuk menghasilkan klasifikasi yang reliable. Ide utamanya adalah membangun version space yang mengandung sekumpulan fungsi hipotesis/classifier. Jika perkiraan dari fungsi hipotesis tersebut adalah benar, maka rule klasifikasi dari version space yang disebut dengan *unnanimous-voting rule,* akan digunakan untuk menjamin bahwa

jika suatu data baru diklasifikasikan, maka data tersebut adalah benar diklasifikasikan.

Dalam tugas akhir ini, SVM ini akan dikombinasikan dengan version space. Fungsi hypotesis atau hyperplane yang dihasilkan SVM akan diuji menggunakan unnanimous voting rule, sehingga hasil klasifikasinya menjadi reliable berdasarkan unanimous-voting rule tersebut. Parameter yang digunakan pada pengujian ini adalah parameter SVM dan parameter kernel yang dipakai oleh SVM itu sediri.

1.2 Perumusan Masalah

Permasalahan yang menjadi objek penelitian dalam tugas akhir ini adalah:

1. Bagaimana menerapkan pre-processing terhadap dokumen web hingga menghasilkan output yang menjadi data masukan yang sesuai untuk mesin pembelajaran SVM.
2. Bagaimana menerapkan metode Support Vector Machine yang dikombinasikan dengan Version Space untuk menghasilkan klasifikasi yang reliable.
3. Bagaimana menentukan parameter-parameter masukan SVM agar menghasilkan version space yang konsisten dengan data training sehingga menghasilkan klasifikasi yang reliable.

Yang menjadi batasan masalah dalam objek penelitian ini adalah:

1. Menggunakan inputan dokumen web berbahasa inggris yang berformat HTML sebagai data set, yaitu dataset web yang tersediakan pada CMU World Wide Knowledge Base (Web - KB) project, 7 sectors.
2. Teks yang diambil dalam dokumen web sebagai feature adalah teks yang ada dalam tag title dan tag meta (metatag description dan metatag keywords).
3. Meskipun SVM dapat menangani kasus multiclass, tetapi pada version space belum dapat menangani kasus multiclass. Sehingga penerapan klasifikasi SVM dengan version space ini hanya menangani kasus binary class (2 kelas).
4. Hanya menggunakan 2 kernel terbaik yang banyak dipakai pada aplikasi SVM yaitu kernel RBF dan kernel Polinomial.

1.3 Tujuan

Berdasarkan rumusan masalah diatas, adapun yang menjadi tujuan penelitian tugas akhir ini adalah :

1. Membangun perangkat lunak yang melakukan pre-processing terhadap dokumen web sehingga bisa mendapatkan suatu dataset untuk machine learning.
2. Menerapkan Support Vector Machine dan Version Space sebagai classifier untuk mendapatkan klasifikasi yang reliable.
3. Melakukan pengujian terhadap beberapa parameter kernel dan parameter SVM untuk mendapatkan performansi yang optimal.

4. Menerapkan feature selection dan melakukan pengujian terhadap sejumlah atribut untuk mendapatkan performansi yang optimal.

1.4 Metodologi Penyelesaian Masalah

Metodologi pembahasan yang digunakan dalam penelitian tugas akhir ini adalah:

1. Studi literatur
   a. Pencarian referensi
      Mencari referensi yang berhubungan dengan Web Mining, Text Classification, Web Page Classification, Machine Learning, Support Vector Machine, Version Space, Reliable Klasifikasi, Feature Selection, Statistical Learning, dan hal-hal lain yang berhubungan dengan judul Tugas Akhir ini, dengan tujuan memberikan gambaran dasar teori yang detail dan jelas.
   b. Pendalaman materi
      Mempelajari dan memahami teknik dan cara kerja Support Vector Machine dan Version Space.
2. Analisa kebutuhan dan perancangan perangkat lunak.
   a. Menentukan kebutuhan sistem seperti identifikasi input/output sistem dan identifikasi hardware/software.
   b. Merancang gambaran umum sistem seperti alur proses sistem dan perancangan kebutuhan program.
3. Implementasi.
   Mengimplementasikan perangkat lunak yang digunakan dalam proses pre-processing, dimana hasil dari pre-processing akan menjadi masukan bagi mesin pembelajaran.
4. Analisa hasil implementasi.
   Menganalisa semua keluaran dari hasil klasifikasi terhadap dokumen web yaitu persentase data yang benar diklasifikasikan oleh classifier.
5. Penyusunan laporan tugas akhir.
   Hasil penelitian akan disusun dalam laporan berupa buku TA yang meliputi aspek-aspek penelitian yaitu teori dan implementasi.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Kesimpulan untuk tugas akhir ini :

1. Untuk kasus data yang dipakai pada eksperimen ini, kekonsistenan classifier terhadap data banyak ditemukan pada saat nilai C semakin besar yaitu C ≥ 10. Hal ini disebabkan oleh karena parameter C mempengaruhi version space. Volume version space semakin besar saat C semakin besar.
2. Akurasi testing semakin menurun ketika nilai C semakin besar yaitu saat C>10. Hal ini disebabkan karena C yang semakin besar membuat volume dari version space juga semakin besar yang membuat data baru yang harus diklasifikasikan kemungkinan berada pada wilayah volume version space yaitu wilayah dimana data tidak diketahui kelasnya sehingga akurasinya menurun.
3. Nilai parameter Polinomial dan RBF yang terlalu besar menyebabkan algoritma sulit untuk menemukan kekonsistenan hyperplane yaitu $\gamma > 0.01$ (untuk domain 1) dan $\gamma > 0.1$ (untuk domain 2), $p > 5$ (untuk domain 1 dan 2)
4. Terjadi overfitting saat nilai $\gamma \geq 0.005$ (untuk domain 1) dan $\gamma \geq 0.01$ (untuk domain 2), $p \geq 0.5$ (untuk domain 1) dan $p > 5$ (untuk domain 2)
5. Feature selection sangat baik digunakan dalam pengklasifikasian dokumen web ini karena mampu menghasilkan akurasi testing yang maksimal ketika atribut dikurangi, yaitu 82.35 % untuk kernel Polinomial dan 76% untuk kernel RBF pada domain data pertama dengan hanya menggunakan 60-90 atribut dari total 175 atribut. Dan pada domain data kedua 80 % untuk kernel RBF dan Polinomial dengan hanya menggunakan 30 atribut dari total 214 atribut.

5.2 Saran

Saran untuk pengembangan Tugas Akhir ini adalah :

1. Mengembangkan penelitian ini untuk menangani kasus multiclass.
2. Mengembangkan penelitian ini untuk kasus data non-numeric.

Daftar Pustaka

[1] Arul, P. A., Kranthy, K. R., ***Web Page Categorization based on Document Structure***. International Institute of Information Technology, Gachibowli, India.

[2] Eirinaki, Magdalini., *Web Mining : **A RoadMap***, Athens University of Economics and Business, Dept. of Informatics.

[3] Evgueni, N, Smirnov., ***Separable VSSVM : An Implementation of Version Space Support Vector Machine for Seperable Data***. Maastricht University, Netherlands.

[4] Lin, C. J., Hsu, C. W., and Chang, C. C., 2003. ***A Practical Guide to Support Vector Classification***. Nation Taiwan University, Taipe 106, Taiwan.

[5] Liu, Tao., Liu, Shengping., ***An Evaluation on Feature Selection for Text Clustering***. Peking University, Beijing.

[6] Nugroho, S, A., Witarto, B, A., Handoko, D., ***Support Vector Machine:Teori dan Aplikasinya dalam Bioinformatika***. http://www.ilmukomputer.com.

[7] Pierre, J. M., 2000. ***Practical Issues for Automated Categorization of Web Site.*** Metacode Technologies, San Francisco.

[8] Platt, C, John. ***The Simplified SMO Algorithm***. CS229, Autum 2008. Microsoft Research.

[9] Riboni, D. ***Feature Selection for Web Page Classification***. D.S.I, Degle Studi di Milano, Italy.

[10] Roland, Nilson., Jose, Pena. ***Evaluating Feature Selection for SVMs in High Dimensions***. IFM Computational Biology, Linkoping University.

[11] Sembiring, Krisantus., 2007. ***Penerapan Teknik Support Vector Machine untuk Pendeteksian Intrusi pada Jaringan***. ITB.

[12] Smirnov,E,N., Nalbantov, G, I., Sprinkhuizen, I, G., ***Reliable Instance Classification with Version Spaces***. Erasmus University Rotterdam, Netherlands.

[13] Smirnov,E,N., Nalbantov, G, I. Sprinkhuizen, I, G., ***Unanimous Voting using Support Vector Machine***. Erasmus University Rotterdam, Netherlands.

[14] Smirnov,E,N., Nalbantov, G, I., Sprinkhuizen, I, G., ***Version Space Support Vector Machine***. Erasmus University Rotterdam, Netherlands.

[15] Staelin, Carl., ***Feature Selection***. Information Retrieval and Digital Libraries.

[16] Vanderlooy, Stijn., ***Co-Training of Version Space Support Vector Machine***. Transnation University Limburg.

[17] Wang, Lipo., ***Support Vector Machine : Theory and Applications***. Nanyang Technological University.

[18] Niklas, L., Paul, D. ***Quantifying The Impact of Learning Algorithm Parameter Tuning***. Blekinge Institute of Technology, Ronneby, Sweden.