ANALISIS DAN IMPLEMENTASI ALGORITMA ADTBOOST.MH UNTUK KLASIFIKASI DATA YANG MULTI-LABEL

Agung Basuki[1], Yanuar Firdaus A.w.[2], Moch Arif Bijaksana[3]

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

Abstrak

Klasifikasi adalah proses mengelompokkan data-data ke dalam kelas-kelasnya. Pada periode ini penerapan metode klasifikasi yaitu klasifikasi multi-label semakin banyak dibutuhkan dalam aplikasi-aplikasi modern, seperti klasifikasi fungsi protein, gen, dan scene semantik. Masalah yang terjadi adalah bagaimana memprediksi kelas-kelas dari data yang multi-label dengan karakteristik data yang berbeda-beda.

Metode atau algoritma klasifikasi yang diimplementasikan adalah ADTBoost.MH yang merupakan perpaduan ide dari metode boosting dan metode decision tree, dimana dengan memanfaatkan iterasi dalam boosting algoritma ini membangun suatu pengklasifikasi yang disebut dengan Alternating Decision Tree (ADTree). ADTBoost.MH menyelesaikan masalah klasifikasi data multi-label secara binary classification.

Melalui percobaan yang telah dilakukan, algoritma ADTBoost.MH berhasil menunjukkan performansi yang bagus dalam menangani masalah klasifikasi yang dihadapi pada karakteristik data multi-label yang berbeda-beda. Yang lebih utama adalah ADTBoost.MH ini menunjukkan performansi yang lebih bagus untuk mengklasifikasikan data yang multi-label bila dibandingkan dengan algoritma klasifikasi single-label.

Kata Kunci : multi-label, ADTBoost.MH, ADTree, boosting, pengklasifikasi, matriks evaluasi

Abstract

Classification is a process that classify data into their true class. In this periode, multi-label classification methods are increasingly required modern applications, such as protein function classification, gen classification, and semantic scene classification. The problem that arise is how to predict the true class from many different characteristics of multi-label data.

Method or algorithm that implemented is called by ADTBoost.MH algorithm. This algorithm came from the idea of boosting methods and decision tree methods. Using the boosting iteration, this algorithm build an Alternating Decision Tree (ADTree) as a classifier. ADTBoost.MH solving multi-label classification problem with binary classification procedure.

Experiments show that ADTBoost.MH is reliable to solve classification problem from different characteristics of multi-label data. More importantly, this algorithm shows a better performance to classify multi-label data rather than use single-label classification algorithm to classify multi-label data.

Keywords : multi-label, ADTBoost.MH, ADTree, boosting, classifier, evaluation metrics

1. Pendahuluan

1.1 Latar Belakang Masalah

Klasifikasi merupakan salah satu *task* dalam *data mining* yang sudah cukup dikenal, banyak penelitian dan juga pengembangan dilakukan dalam bidang tersebut. Pada periode sebelumnya, penelitian tentang klasifikasi data banyak terfokus dalam kasus *binary classification* (klasifikasi dengan jumlah kelas dua) atau *multiclass classification* (klasifikasi dengan jumlah kelas lebih dari dua) dimana masing-masing *instance* dikategorikan ke dalam satu kelas (label) tunggal atau disebut juga dengan klasifikasi *single-label*. Namun seiring dengan meluasnya bidang aplikasi untuk *task* data mining (dalam hal ini klasifikasi data), terutama didorong oleh masalah klasifikasi dokumen, bioinformatik dan diagnosis-diagnosis medikal, muncul suatu masalah baru dimana suatu *instance* tidak hanya menjadi anggota satu kelas saja, melainkan merupakan anggota dari beberapa kelas.

Keanggotaan yang tidak *mutually exclusive* tersebut kemudian disebut sebagai kasus *multi-label*. *Multi-label* selalu terkait dengan data yang ambigu, dimana setiap satu *intance* merupakan anggota dari sejumlah kelas (label) yang berbeda. Hal tersebut tentunya menambah tingkat kesulitan dalam memprediksi kelas-kelas dari suatu data. Permasalahan tersebut banyak ditemukan dalam kondisi aplikasi saat ini, contohnya pada aplikasi modern seperti klasifikasi fungsi protein, kategorisasi musik, dan klasifikasi *scene* semantik [11]. Untuk menyelesaikan masalah klasifikasi *multi-label* tersebut, dibutuhkan suatu metode pemecahan yang baru, karena tidak semua metode klasifikasi yang telah ada bisa diterapkan. Hal tersebut mendorong suatu penelitian dan pengembangan baru untuk mendapatkan metode klasifikasi yang efektif. Meskipun belum cukup banyak penelitian yang dilakukan, sudah ada beberapa metode baru yang merupakan pengembangan dari metode yang ada untuk menyelesaikan masalah klasifikasi *multi-label* [1, 2, 10].

Di antara metode klasifikasi yang telah ada, *boosting* dianggap sebagai salah satu metode klasifikasi yang cukup handal dan mampu menghasilkan tingkat akurasi yang lebih baik daripada metode klasifikasi lain [10]. Metode tersebut menggabungkan semua *rule* sederhana yang dihasilkan pada setiap iterasinya menjadi satu rule dengan tingkat akurasi yang tinggi. *Boosting* juga mengatur bobot (distribusi) pada data *training* yang bertujuan agar model yang akan dibentuk lebih berkonsentrasi pada data yang sulit diklasifikasikan. Akan tetapi, seperti yang dikatakan oleh Freund dan Mason serta peneliti lainnya, *boosting* kadang menghasilkan model klasifikasi yang susah untuk diinterpretasi dan dimengerti [2, 4]. Oleh karena itu, diterapkan suatu algoritma klasifikasi yang merupakan kombinasi antara teknik *boosting* dan *decision tree* (pohon keputusan) yang disebut *ADTBoost.MH*. Pertama-tama data *multi-label* ditransfomasikan menjadi data biner, kemudian melalui iterasi *boosting* dibangun *ADTree* (*Alternating Decision Tree*) dengan strategi *top-down*. *ADTree* adalah suatu pohon yang terdiri dari node-node pembagi (*splitter*) dan node-node prediksi yang berisi nilai-nilai riil dan menunjukkan keanggotaan suatu data terhadap suatu

kelas. Melalui pembentukan pohon inilah maka model klasifikasi yang dihasilkan lebih mudah untuk diinterpretasikan [2].

Tugas akhir ini akan mengimplementasikan algoritma *ADTBoost.MH* tersebut untuk menyelesaikan masalah klasifikasi data yang *multi-label,* di antaranya untuk data *yeast, scene,* dan *genbase* [13]. Pengklasifikasi yang dihasilkan melalui proses pembelajaran pada data *training* akan diujikan untuk memprediksikan label-label dari sekumpulan data yang disebut data *testing*. Kemudian melalui serangkaian percobaan akan dianalisis kelebihan dan kekurangan dari algoritma tersebut.

1.2 Perumusan Masalah

Dengan mengacu latar belakang di atas, maka permasalahan yang dibahas dan diteliti adalah :

1. Bagaimana melakukan klasifikasi pada data *multi-label* menggunakan algoritma ADTBoost.MH.
2. Bagaimana analisis perbandingan algoritma ADTBoost.MH dengan algoritma lain, jika dilihat dari perhitungan matriks evaluasi untuk masing-masing algoritma.

Sedangkan ruang lingkup yang menjadi batasan dari pembahasan tugas akhir ini adalah :

1. *Dataset* yang digunakan merupakan data *multi-label*.
2. Data yang digunakan untuk analisis merupakan data yang *supervised* (memiliki *class label*).
3. Tidak menangani data *imbalance* dan data *preprocessing*, data *training* dan data *testing* bersih dari *noise*.
4. Input sistem berupa data dengan format sesuai dengan yang telah ditentukan.
5. *Dataset* yang akan digunakan untuk proses analisis adalah data *yeast, scene,* dan *genbase.*
6. Matriks evaluasi dihitung berdasarkan nilai *accuracy, coverage, hamming loss, one-error,* dan *ranking loss*. *Accuracy* menunjukkan keakuratan dari algoritma yang digunakan dalam memprediksi kelas-kelas dari data yang diujikan (data *testing*). Karena alasan itulah, *accuracy* dijadikan sebagai kriteria evaluasi utama.

1.3 Tujuan

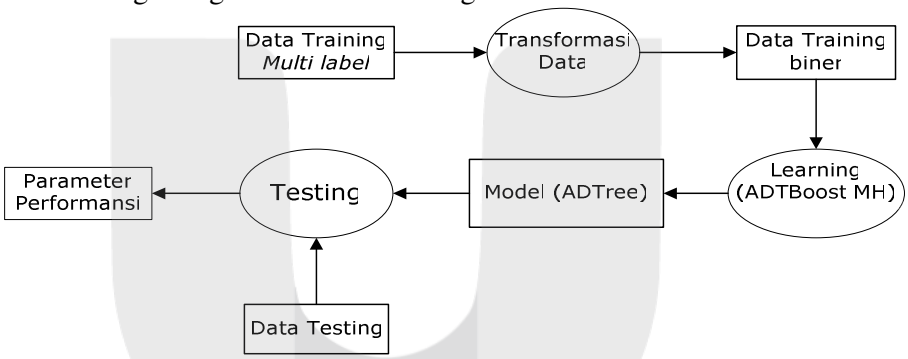
Berdasarkan pada masalah yang telah didefinisikan di atas, maka tujuan tugas akhir ini adalah :

1. Mengimplementasikan algoritma ADTBoost.MH untuk melakukan klasifikasi data yang *multi-label.*
2. Menganalisis keakuratan dan kecepatan algoritma ADTBoost.MH untuk proses klasifikasi data *multi-label*. Parameter akurasi yang digunakan adalah *accuracy, coverage, hamming loss, one-error,* dan *ranking loss*.
3. Membandingkan akurasi dari algoritma ADTBoost.MH dengan algoritma lain menggunakan *tools* melalui analisis perbandingan nilai *accuracy, coverage, hamming loss, one-error, dan ranking loss*.

1.4 Metodologi Penyelesaian Masalah

Pendekatan sistematis/metodologi yang akan digunakan dalam merealisasikan tujuan dan pemecahan masalah di atas adalah dengan menggunakan langkah-langkah sebagai berikut :

1. Studi literatur
   Pencarian, pengumpulan, dan pembelajaran literatur-literatur berupa buku, jurnal, artikel, dan sumber lain yang berhubungan dengan topik tugas akhir ini. Selain itu, dilakukan juga konsultasi kepada pihak-pihak yang menguasai materi.
2. Pengumpulan data
   Mengumpulkan *dataset* yang diperlukan, kemudian memahami maksud dari data-data maupun atribut tersebut serta melakukan *preprocessing* terhadap data sehingga data siap untuk digunakan.
3. Analisis kebutuhan dan perancangan perangkat lunak
   Menspesifikasikan kebutuhan-kebutuhan perangkat lunak kemudian membangun rancangan dari perangkat lunak tersebut. Adapun sistem yang akan dibangun digambarkan dalam bagan di balik ini



Gambar 1-1: Diagram Proses Klasifikasi.

Sebelum proses learning (pembangunan model) sebelumnya dilakukan transformasi terhadap data yang *multi-label.* Proses tersebut mengubah satu *instance* menjadi k *instance* yang diberi label biner, dimana k merupakan jumlah kelas yang ada. Setelah diperoleh model hasil *learning* oleh algoritma ADTBoost.MH maka kemudian dilakukan testing. Keluaran dari proses testing tersebut adalah kecepatan dan nilai-nilai performansi dari algoritma yang digunakan sesuai dengan parameter pengukuran yang ditetapkan (*accuracy, coverage, hamming loss, one-error,* dan *ranking loss*).

4. Implementasi
   Implementasi rancangan dengan membangun suatu perangkat lunak untuk melakukan klasifikasi objek-objek ke dalam label tertentu.
5. Pengujian sistem dan analisis hasil
   Melakukan pengujian dengan data *training* sebagai data pembangun model dan data *testing* untuk mengukur performansi dan akurasi perangkat lunak, parameter yang digunakan adalah *accuracy, coverage, hamming loss, one-error, dan ranking loss*. Menganalisis pengaruh jumlah iterasi *boosting* yang dilakukan terhadap akurasi dan kecepatan pembuatan model yang dihasilkan. Menganalisis pengaruh jumlah kelas terhadap akurasi yang dihasilkan.
6. Pengambilan kesimpulan dan penyempurnaan laporan Tugas Akhir.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan percobaan dan analisis yang telah dibahas dan dilaksanakan pada bab 4, maka dapat disimpulkan beberapa hal sebagai berikut :

1. Algoritma *ADTBoost.MH* dapat diterapkan untuk menyelesaikan masalah klasifikasi data *multi-label*, baik untuk data yang bertipe *numerik* ataupun data yang bertipe diskrit.
2. Performansi yang lebih bagus didapatkan ketika algoritma *ADTBoost*.MH diterapkan pada data yang bukan *numerik* (diskrit). Terbukti dalam pengujian pada sub bab 4.3.1.1 dimana untuk *dataset genbase* yang bertipe diskrit menghasilkan *accuracy* di atas 80%, sedangkan untuk *dataset yeast* dan *scene* yang bertipe *numerik* hanya menghasilkan kisaran angka *accuracy* antara 50%-60 %.
3. Perubahan jumlah iterasi memberikan pengaruh terhadap performansi algoritma *ADTBoost.MH*, kebutuhan akan jumlah iterasi yang tepat untuk dilakukan tergantung dari karakteristik *dataset* yang akan diklasifikasikan.
4. *ADTBoost.MH* sebagai algoritma klasifikasi *multi-label* memberikan performansi yang lebih baik untuk menangani kasus data yang *multi-label* bila dibandingkan dengan algoritma klasifikasi *single-label*. Hal tersebut disebabkan karena algoritma klasifikasi *single-label* tidak memperhatikan keterhubungan antar label seperti yang dilakukan oleh algoritma *ADTBoost.MH*.
5. Penggunaan label dalam data mempengaruhi performansi dari pengklasifikasi. Semakin sedikit label yang digunakan maka performansi pengklasifikasi semakin naik.

5.2 Saran

Sebagai acuan dalam melengkapi atau memperbaiki hasil analisis data yang dilakukan dalam tugas akhir ini. Ada beberapa saran yang dapat dijadikan pertimbangan bagi analisis data selanjutnya, di antaranya :

1. Data *multi-label* yang digunakan sebaiknya merupakan data yang telah mengalami *preprocessing* dengan teknik *preprocessing multi-label* pula.
2. Dibutuhkan suatu usaha pencarian yang lebih efektif dalam menentukan *rule* baru dalam tiap iterasi algoritma, sehingga mempercepat waktu proses *training*.
3. Perlu diadakan juga pengujian terhadap penentuan *threshold* dalam mengklasifikasikan data untuk meningkatkan performansi dari klasifikasi *multi-label*.
4. Pengembangan aplikasi agar tidak terbatas pada tipe data tertentu saja.

Daftar Pustaka

[1] Clare, A. (2003). *Machine learning and data mining for yeast functional genomics*. Wales, UK : University of Wales Aberystwyth.

[2] Comité, F. D., Gilleron, R. dan Tommasi, M. (2003). "Learning Multi-label Alternating Decision Tree from Texts and Data". In *MLDM*, 35–49.

[3] Fan, Rong-En dan Lin, Chih-Jen . A Study on Threshold Selection for Multi-label Classification.

[4] Freund, Y. dan Mason, L. (1999). "The alternating decison tree learning algorithm". In *Proc. 16th International Conf. On Machine Learning,* 124-133.

[5] Han, Jiawei dan Kamber, Micheline. (2006). *Data Mining : Concepts and Techniques 2nd Edition.* San Fransisco : Morgan Kaufmann Publishers.

[6] Kohavi, R dan Kunz, C. (1997). "Option decision trees with majority votes". In *Machine Learning: Proceedings of the 14th Internatinal Conference,* pages 211-218.

[7] Liu. Kuang-Yu, Lin, J., Zhou, X. dan Wong, S. (2005). "Boosting alternating decision trees modeling of disease trait information". In *BMC Genetics* 2005, 6:S132.

[8] Pramudino, Iko. (2003). "Pengantar Data Mining : Menambang Permata Pengetahuan di Gunung Data". IlmuKomputer.Com.

[9] Schapire, R. E. dan Singer, Y. (1998). "Improved boosting algorithms using confidence-rated predictions". In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT-98),* pages 80-91. New York, July 24-26 1998. ACM Press.

[10] Schapire, R. E. dan Singer, Y. (2000). "BoosTexter: A boosting-based system for text categorization". *Machine Learning*, 39(2/3), 135-168.

[11] Tan, P., Steinbach, M. dan Kumar, V. (2004). *Introduction to Data Mining*. Addison-Wesley.

[12] Tsoumakas, G. dan Katakis, I. (2007). "Multi-Label Classification: An Overview". *International Journal of Data Warehousing and Mining*, 3(3), 1-13.

[13] 2007. *Machine Learning & Knowledge Discovery Group*. http://mlkd.csd.auth.gr/multilabel.html. diakses pada tanggal 15 Desember

2007.

[14] 2007. *Munich Information Center for Protein Sequence*. http://mips.gsf.de. Diakses pada tanggal 15 Desember 2007.