BAB 1

PENDAHULUAN

1.1 Latar belakang masalah

Perkembangan teknologi yang pesat dalam bidang basis data memungkinkan orang, kelompok, atau organisasi untuk mengumpulkan dan menyimpan data dalam jumlah besar yang berasal dari aktivitas sehari-hari mereka. Data tersebut tidak cukup hanya dikumpulkan dan disimpan saja, agar memberikan manfaat maka masih perlu diolah dengan suatu mekanisme yang disebut data mining. Data mining adalah proses menemukan pengetahuan yang berguna secara otomatis dari tempat penyimpanan data yang besar. Data mining sering juga dianggap sebagai bagian tak terpisahkan dari *Knowledge Discovery in Databases* (KDD). Sedangkan KDD sendiri adalah keseluruhan proses mengubah *raw* data menjadi *knowledge* yang berguna.

Salah satu task data mining yang menjadi inspirasi dari Tugas Akhir ini adalah klasifikasi. Klasifikasi terbagi dalam single classification model (misal: *Decision tree*, *Rule-based Classifier, Nearest-neighbor Classifier, Bayesian Classifier, Artificial Neural Network, Support Vector Machine*) dan *multiple classification model/ensemble method* (misal: *Bagging, Boosting, Random Forest*) [2]. *Ensemble method* mampu memperbaiki akurasi dari *single classification model* yaitu dengan melakukan kombinasi terhadap prediksi yang dibuat oleh beberapa *classifier* [2]. Fokus Tugas Akhir ini adalah *Boosting* dengan menerapkan algoritma Adaboost.

Boosting adalah sebuah prosedur iterative untuk mengubah data training secara adaptive dengan fokus pada data yang sulit diklasifikasikan [2]. Tugas-tugas dalam data mining sering kali melibatkan data dalam jumlah sangat besar dan berdasarkan uraian tentang *boosting* di atas, maka kita harus menyediakan main memory dengan kapasitas memadai untuk menampung data training baru, yang mana ukurannya sama dengan data training asli. Jika kapasitas main memory kurang maka proses *boosting* menjadi lambat oleh karena terjadinya swapping dari main memory ke harddisk. Solusi dengan up-grade main memory bukanlah yang terbaik karena main memory yang tersedia di pasar masih memiliki kapasitas yang terbatas sedangkan data bisa

berlipat-lipat ukurannya seiring waktu, selain itu harga main memory yang masih mahal juga merupakan kendala tersendiri bagi kalangan tertentu. Masalah seperti inilah yang dicoba untuk ditemukan solusinya dalam Tugas Akhir ini dengan menerapkan pendekatan partisi data training pada proses *boosting* [1].

1.2 Perumusan masalah

Berdasarkan uraian sebelumnya maka diperlukan solusi alternatif selain meng-upgrade main memory. Solusi alternatif yang layak untuk dipertimbangkan adalah menerapkan *boosting*. Namun *boosting* di sini berbeda dari *boosting* secara normal karena terlebih dahulu dilakukan partisi terhadap data training asli sehingga menghasilkan beberapa bagian. Kemudian *boosting* diterapkan pada masing-masing bagian tersebut. Hasil dari proses *boosting* pada masing-masing bagian kemudian dikombinasikan untuk memperoleh prediksi final terhadap data training. Dalam rangka mewujudkan solusi alternatif ini maka dapat dirumuskan beberapa masalah sebagai berikut :

1. Algoritma apakah yang akan dipakai dalam proses *boosting*?
2. Teknik apakah yang dipakai untuk tahap pembentukan *classifier* pada setiap iterasi dan partisi (misal : teknik *decision tree* induction, rule-based *classifier*, atau neural network).?
3. Bagaimana cara untuk mempartisi data training asli?
4. Data training sebaiknya dibagi dalam berapa partisi?
5. Bagaimana cara untuk mengkombinasikan semua *classifier* yang diperoleh dari setiap partisi sehingga menghasilkan *classifier* final?
6. Spesifikasi hardware dan data yang seperti apakah yang mampu merepresentasikan kondisi seperti disebutkan pada latar belakang masalah?
7. Bagaimana cara melakukan pengujian dan analisis hasil *boosting* yang seperti ini?
8. Bagaimanakah performansi *boosting* dengan partisi data training jika dibandingkan *boosting* yang biasa?

Dalam Tugas Akhir ini ada beberapa batasan masalah yang perlu diperhatikan, antara lain :

1. *Boosting* pada Tugas Akhir ini menggunakan algoritma Adaboost.
2. Teknik *decision tree* induction dipakai dalam tahap pembentukan *classifier* pada setiap partisi.

Hipotesa awal dari tugas akhir yang hendak saya buat ini adalah klasifikasi menggunakan *boosting* dengan pendekatan partisi terhadap data training memiliki hasil akhir sebaik *boosting* yang biasa dengan beberapa keunggulan (diantaranya : tidak boros memori).

1.3 Tujuan

Tujuan penulisan Tugas Akhir ini adalah :

1. Membangun perangkat lunak yang mengimplementasikan teknik *boosting* untuk menyelesaikan masalah klasifikasi dalam data mining.
2. Menganalisis pengaruh partisi data training terhadap proses *boosting*.
3. Menganalisis kelayakan dari pendekatan partisi data training pada proses *boosting* menggunakan algoritma Adaboost.
4. membuktikan apakah keterbatasan *hardware* dalam data mining dapat diatasi dengan teknik *boosting* berdasarkan pendekatan partisi terhadap *data training*.

1.4 Metodologi penyelesaian masalah

Metode penyelesaian masalah dalam penulisan tugas Akhir ini disusun dalam langkah-langkah sebagai berikut:

1. Studi literatur, yang dilakukan dengan membaca dan mempelajari beberapa sumber tertulis (makalah, buku dan jurnal) yang berkaitan dengan data mining, klasifikasi, *ensemble method*, *boosting*, algoritma Adaboost, pendekatan partisi data training terhadap algoritma Adaboost, *decision tree* induction
2. Pengumpulan dan analisis data yang mendukung implementasi dan analisis algoritma Adaboost dalam perangkat, yaitu data dari kompetisi PAKDD 2006
3. Analisis kebutuhan dan perancangan perangkat lunak, untuk menentukan kebutuhan pembangan pembangunan perangkat lunak, serta perancangan struktur data dan aktivitas perangkat lunak yang dibangun.
4. Implementasi, yang merupakan langkah penerapan rancangan yang telah dibuat ke dalam perangkat lunak yang dapat digunakan untuk menyelesaikan masalah klasifikasi dengan algoritma Adaboost menggunakan pendekatan partisi data training.
5. Pengujian dan analisis hasil, yaitu langkah untuk menilai performansi algoritma Adaboost dengan pendekatan partisi data training dalam perangkat lunak terhadap algoritma Adaboost tanpa pendekatan partisi data training.