Analisis Perbandingan CPU dan GPU (CUDA) Pada Klasifikasi Data Mining dengan Menggunakan Metode K-Nearest Neighbor Kernel Algorthm

Faris Muhammad[1], Ibnu Asror, S.T.,MT [2],Indra Lukmana Sardi. S.T.,MT [3],

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]farismuhammad@students.telkomuniversity.ac.id, [2]iasror@telkomuniversity.ac.id,
[3]indraluk@telkomuniversity.ac.id

Abstrak

Data mining merupakan proses semi-otomatis untuk pengekplorasian data yang berjumlah besar gunanya untuk mendapatkan pola yang berguna. Data mining ini merupakan proses gabungan antar bidang-bidang terutama adalah machine learning, analisis statistik dan basis data. Data mining berusaha untuk menemukan kaidah dan pola dari data.Salah satu task yang penting dalam data mining adalah classification (klasifikasi). Klasifikasi ini dapat dideskripsikan sebagai berikut: terdiri dari data *input* yang disebut juga sebagai *training set* terdiri dari sejumlah *examples* (record) yang masing-masing memilki sejumlah atribut atau disebut juga fitur. Adapun tujuan klasifikasi ini adalah untuk menganalisa data *input* dan mengembangkan sebuah model yang akurat untuk setiap kelas berdasarkan beberapa *variabel prediktor*. Untuk menghasilkan informasi saat melakukan proses data *mining* kendala yang dihadapi adalah banyaknya jumlah data sehingga proses yang dilakukan oleh CPU akan berjalan sangat lambat apabila dirasakan. Untuk menanggulangi masalah ini maka proses data mining menggunakan GPU menjadi salah satu solusi dalam menangani *running time* yang lambat dan akurasi yang kurang baik. Melalu tugas akhir ini penulis akan mencoba menganalisis sebuah algoritma KNN Kernel, Metode ini merupakan perkembangan dari metode KNN *Standard.* Dimana pada metode KNN *Standard* proses klasifikasi dilakukan dengan melihat sejumlah k tetangga terdekat, dan akan diklasifikasikan berdasarkan jumlah kelas terbanyak pada sejumlah k tetangga terdekatnya. *Classifier* tersebut diuji menggunakan 3 fungsi Kernel. Hasil yang didapat dari percobaan penulis yaitu pada pembagian *5 fold* total waktu CPU1: 1,68 s,CPU2: 15,63 s,GPU1: 12,29 s,GPU2: 4,61 s. dan pada pembagian *10 fold* total waktu CPU1: 1,53 s,CPU2: 15,27 s,GPU1: 12,05 s,GPU2: 4,55. Akurasi yang didapatkan pada pembagian *5 fold* 63,87% dan pembagian *10 fold* 64,30% pada semua perangkat.

Kata Kunci : data mining,klasifikasi,CPU, GPU,KNN Kernel

Abstract

Data mining is a semi-automatic process for exploring and analyzing large amounts of data to get useful patterns. Data mining is a joint process between fields, especially machine learning, statistical analysis and database. Data mining tries to find the rules and patterns of data. One important task in data mining is classification (classification). This classification can be described as follows: consists of input data which is also called training set consisting of a number of examples (records) which each have a number of attributes or also called features. The purpose of this classification is to analyze input data and develop an accurate model for each class based on several predictor variables. To produce information when doing data mining process, the obstacles faced are the large amount of data so that the process carried out by the CPU will run very slowly when felt. To overcome this problem, the data mining process uses GPU to be one of the solutions in handling slow running time and poor accuracy. Through this final project the author will try to analyze a KNN Kernel algorithm, this method is a development of the KNN Standard method. Where in the KNN Standard method the classification process is carried out by looking at a number of the closest neighbors, and will be classified based on the number of classes in the number of the closest neighbors. The classifier is tested using 3 Kernel functions. The results obtained from the authors' experiments are that the division of 5 fold total CPU time1: 1.68 s, CPU2: 15.63 s, GPU1: 12.29 s, GPU2: 4.61 s. and in dividing the 10 fold total CPU time1: 1.53 s, CPU2: 15.27 s, GPU1: 12.05 s, GPU2: 4.55. Accuracy obtained at 5 fold division is 63.87% and division of 10 fold is 64.30% on all devices.

Keywords: data mining,classication,CPU,GPU,KNN Kernel

1. Pendahuluan

1.1 Latar Belakang

Seiring dengan perkembangan teknologi dalam hal pengumpulan data dan penyimpanan data dapat menyebabkan tumpukan data yang banyak. Dengan adanya kumpulan data yang banyak, maka timbulah suatu kebutuhan untuk bisa memanfaatkan data tersebut. Pemanfaatan data tersebut tentunya bertujuan untuk mendapatkan informasi penting dari pola-pola data yang terbentuk. Proses untuk mendapatkan informasi atau pola-pola berharga dari sekumpulan data tersebut lah dinamakan Data mining[1]. Klasifikasi merupakan suatu metode dari data mining. Ini merupakan metode prediktif yang melakukan pembelajaran terhadap data-data yang sudah

ada sehingga menghasilkan suatu model yang digunakan untuk memprediksi data-data baru. Salah satu alogritma klasifikasi yang terkenal adalah K-nearest neighbor (KNN)[3].

K-nerest neighbor (KNN) adalah suatu metode yang menggunakan alogritma supervised dimana hasil query instance yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada KNN. Tujuan dari algoritma ini adalah mengklasifikasikan obyek baru berdasarkan atribut dan traning sample[3]. Algoritma KNN sangatlah sederhana, berkerja berdasarkan jarak terpendek dari test sample dan training sample. Penulis melanjutkan penelitian sebelumnya yang hanya menggunakan KNN dalam pemrosesan *data mining.* Namun, pendekatan KNN berdasarkan jarak saja mempunyai akurasi yang cukup kecil sehingga diterapkan pembobotan terhadap jarak tersebut menggunakan Kernel[12][13].

Pada proses data mining kendala berupa banyaknya jumlah data sehingga menyebabkan running time untuk melakukan analisis berjalan sangat lambat sering dialami. Belum lagi akurasi yang kurang tinggi menyebabkan proses analisis kurang sempurna dan mendapatkan kredibilitas rendah. Hal ini biasa disebabkan oleh kemampuan CPU dalam melakukan proses data masih sangat terbatas. Untuk menanggulangi hal ini maka dilakukan lah pemprosesan data pada GPU dimana pemprosesan data di GPU akan meningkatkan running time dan akurasi dari sebuah data itu sendiri[11][12][13].Pada karya tulis ini penulis akan menggunakan algoritma K-Nearest neighbor Kernel untuk melakukan perbandingan pemprosesan data pada CPU dan GPU menggunakan Dataset Diabetes Retinopathy.

1.2 Topik dan Batasannya

Perumusan masalah dalam tugas akhir ini dibagi kedalam beberapa poin, yaitu:

1.Bagaimana cara melakukan proses klasifikasi *data mining* pada CPU dan GPU?

2.Bagaimana cara algoritma *K-Nearest Neighbor Kernel* dapat melakukan pemrosesan data melalui GPU ?

3. Bagaimana hasil *running time* dan *accuracy* yang terjadi di CPU dibandingkan dengan di GPU?

Adapun batasan masalah pada tugas akhir ini adalah sebagai berikut:

1. Menggunakan *K-nearest neighbor kernel Algorithym.*
2. Pemprosesan data diharuskan menggunakan aplikasi CPU dan GPU.
3. Pemprosesan data difokuskan pada peningkatan akurasi dan *running time*.
4. Hardware yang digunakan harus menggunakan *NVIDIA Graphic card*.

1.3 Tujuan

Tujuan dari penulisan tugas akhir ini adalah:

1. Melakukan integrasi sistem agar dapat menjalankan aplikasi CUDA yang memiliki fungsi untuk menjalankan pemorsesan data pada GPU.
2. Melakukan pemrosesan data menggunakan *K-nearest neighbor Kernel Algorithm* dengan menggunakan CPU dan GPU.
3. Menganalisis perbandingan pemprosesan data yang dilakukan di CPU dan GPU.

1.4 Organisasi Tulisan

Adapun bagian-bagian selanjutnya pada TA ini adalah :

1. Landasan Teori

   Pada bagian ini menjelaskan apa saja teori yang mendukung dengan topik TA yang dikerjakan seperti pengertian Data Mining, Pengertian KNN-Kernel dan lainnya.

2. Perancangan Sistem

   Pada bagian ini penulis menjelaskan bagaimana alur jalannya program penulis dari awal pemrosesan data hingga data mining tersebut berhasil diproses.

3. Pengujian dan Analsis

   Pada bagian ini penulis menjelaskan bagaimana pengujian dari pemrosesan data ini dan akan menganalisis hasil yang ada dari hasil pengujian yang penulis buat.

4. Kesimpulan dan Saran

   Pada bagian ini penulis memberi kesimpulan mengenai hasil yang penulis teliti dan memberi saran atas hasil yang penulis dapatkan

2. Studi Terkait

2.1 Definisi Data Mining

Data mining adalah suatu istilah yang digunakan untuk menemukan pengetahuan yang tersembunyi di dalam *database*. Data mining merupakan proses semi otomatik yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstrasi dan mengidentifikasi informasi pengetahuan potensial dan juga bermafaat yang tersimpan didalam database besar[1].

Data *mining* adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam *database*, data warehouse atau penyimpanan informasi lainnya. Data *mining* berkaitan dengan bidang ilmu – ilmu lain seperti, database system, ata warehousing, statistik, *maching learning* dan sebagainya[1].

Data mining didefinisikan sebagai proses menemukan pola-pola di dalam data. Proses ini otomatis atau seringnya semiotomatis. Pola yang ditemukan harus penuh arti dan pola tersebut memberikan keuntungan , biasanya keuntungan secara ekonomi. Data yang dibutuhkanpun dalam jumlah besar[2].