Jurnal Teknologi Maklumat & Multimedia **2**(2005): 1-17

Perlombongan Data Prestasi Pelajar Siswazah Menggunakan Kaedah Aruhan Berasaskan Atribut

SITI ROHAIDAH AHMAD & AZURALIZA ABU BAKAR

ABSTRAK

Kertas kerja ini membincangkan dengan terperinci kaedah aruhan berasaskan atribut yang merupakan satu kaedah perlombongan data. Pelbagai jenis pengetahuan boleh diperoleh melalui kaedah ini antaranya petua pengelasan, petua ciri, petua pengelasan kuantitatif, petua ciri kuantitatif dan sebagainya. Konsep pengitlakan dan ringkasan merupakan perkara asas dalam melaksanakan kaedah ini. Pengitlakan data dilaksanakan ke atas satu set data yang relevan dengan cara penghapusan atribut, pohon konsep menaik, mengawal proses pengitlakan dengan menetapkan nilai ambang bagi atribut, pembilang perambatan dan nilai fungsi jumlah yang lain. Dua algoritma telah dibina iaitu algoritma kaedah aruhan berasaskan atribut dan algoritma arahan bahasa pertanyaan piawai. Algoritma arahan bahasa pertanyaan piawai bertindak sebagai perantaraan dengan pangkalan data dalam melaksanakan segala arahan bahasa pertanyaan piawai dengan pangkalan data hubungan. Kedua-dua algoritma ini saling berkaitan dalam melaksanakan kaedah aruhan berasaskan atribut yang berorientasikan bahasa pertanyaan piawai. Satu pangkalan data hubungan telah direka bentuk dan dibangunkan untuk menyimpan data pelajar siswazah Fakulti Teknologi dan Sains Maklumat, UKM. Data ini digunakan dalam menguji kedua-dua algoritma tersebut. Satu set petua pengelasan dihasilkan yang mengandungi pengetahuan berkenaan pencapaian pelajar siswazah. Petua-petua ini diuji menggunakan set data uji untuk menentukan ketepatan pengelasannya. Hasil uji kaji menunjukkan set petua yang dihasilkan boleh digunakan untuk mengelaskan pencapaian pelajar siswazah pada masa akan datang.

ABSTRACT

This paper discusses in detail an attribute-oriented induction technique which is one of the data mining techniques. Many types of knowledge can be discovered through this concept such as classification rule, characteristic rule, quantitative classification rule and quantitative characteristic rule. In the implementation, generalization and summarization are the two fundamentals of concepts involved. Data generalization is executed into a set of

relevant data by using elimination attribute procedure, concept-tree climbing, controlling generalization process by determining the threshold value for the attribute, propagation of counts, and other amount of value functions. Two algorithms are developed, which are attribute-oriented induction and standard query language instruction. The standard query language instruction algorithm acts as a medium for the database to execute standard query language instruction to the database link. Both are related to each other in the execution of attribute-oriented induction which is based on standard query language. A database link design is developed to store all data of the postgraduate students of the Faculty of Information Science and Technology, UKM, which was tested using both algorithms. A set of classification rules containing knowledge on the performance of the postgraduate students is obtained. These rules are then tested with new data of those students in order to determine the accuracy of the classification rule. The experimental results show that the rules obtained can be used in the future to determine the performance of postgraduate students.

PENGENALAN

Jumlah maklumat dalam dunia ini semakin hari semakin bertambah. Oleh itu, satu kaedah diperlukan untuk menganalisis maklumat bagi mencari hubungan yang tersembunyi di antara maklumat yang boleh mewakili sejumlah maklumat yang besar tersimpan dalam pangkalan data. Penemuan pengetahuan dalam pangkalan data atau perlombongan data merupakan salah satu isu penting dalam pembangunan data dan sistem yang berasaskan pengetahuan. Pertumbuhan saiz dan pangkalan data yang sedia ada jauh melebihi keupayaan manusia untuk menganalisis data (Cheung et al. 2000), tambahan pula tugas untuk mengekstrak pengetahuan daripada pangkalan data adalah di luar kemampuan manusia melakukannya secara efisyen dan cekap. Sebagai contoh, pangkalan data *Wal_Mart* mengumpul sebanyak 20 juta transaksi setiap hari (Fu 1996).

Melalui pengekstrakan pengetahuan daripada pangkalan data, pangkalan data akan menjadi kaya dengan maklumat yang berguna serta sumber maklumat yang dipercayai untuk capaian pengetahuan dan pengesahan (Fu 1996). Penemuan pengetahuan boleh digunakan dalam pengurusan maklumat, proses membuat keputusan, proses kawalan dan lain-lain aplikasi (Fu 1996). Oleh itu perlombongan data atau penemuan pengetahuan menjadi perkara penting dan mencabar dalam bidang penyelidikan (Silberschatz et al. 1991; Silberschatz et al. 1996). Penyelidikan dalam pelbagai bidang seperti sistem pangkalan data, sistem berasaskan pengetahuan, kepintaran buatan, pembelajaran mesin, perolehan pengetahuan, statistik, pangkalan data ruang (*spatial*) dan visualisasi data telah mula menunjukkan minat terhadap perlombongan data (Chen et al. 1996).

Penemuan pengetahuan daripada pangkalan data ditakrifkan sebagai menggambarkan pengetahuan dalam pangkalan data, melibatkan pengekstrakan data, pencarian data, penjelajahan data, pemprosesan bentuk data, penggalian data dan penuaian maklumat (Turban & Aronson 2001). Pembelajaran adalah salah satu ciri penting dalam diri manusia dan kepintaran mesin, sebabnya pangkalan data hubungan digunakan dengan meluas dalam banyak aplikasi dan ini merupakan satu kelebihan untuk mempelajari pengetahuan dalam bentuk petua daripada data yang tersimpan dalam pangkalan data. Banyak teori, algoritma dan sistem yang masih dalam kajian dibina untuk pembelajaran mesin. Sistem pangkalan data hubungan digunakan dengan meluas dalam pelbagai aplikasi. Melalui pembelajaran daripada pangkalan data, petua pengetahuan boleh diekstrak daripada jumlah data yang besar dan hubungan yang menarik antara data boleh ditemui secara automatik. Tambahan pula, sistem pangkalan data hubungan menyediakan banyak ciri-ciri yang menarik untuk pembelajaran mesin. Pangkalan data hubungan menyimpan maklumat secara berstruktur dan tersusun. Setiap rekod dalam pangkalan data boleh digambarkan sebagai formula logikal iaitu dalam bentuk konjungsi normal (Cai et al. 1991).

Terdapat dua jenis petua yang boleh diperolehi daripada pangkalan data hubungan iaitu petua ciri dan petua pengelasan. Petua ciri adalah pernyataan ciri-ciri data yang disimpan dalam pangkalan data hubungan. Sebagai contoh, simptom mengenai sejenis penyakit boleh diringkaskan sebagai petua ciri. Manakala, petua pengelasan pula adalah satu pernyataan mengenai perbezaan satu kelas dengan kelas yang berlainan. Sebagai contoh, bagi membezakan satu penyakit dengan penyakit yang lain, petua pengelasan seharusnya meringkaskan simptom perbezaan penyakit tersebut dengan penyakit-penyakit yang lain (Cai et al. 1991).

Aplikasi perlombongan data yang besar melibatkan proses membuat keputusan yang besar di mana ia boleh mencapai berbilion bait data. Oleh itu, kecekapan aplikasi adalah amat penting. Pengelasan merupakan fungsi utama dalam perlombongan data di mana rekod dalam pangkalan data dianggap sebagai sampel data yang akan dianalisis secara tersusun untuk menghasilkan model pengelasan (Fayyad et al. 1996a; Fayyad et al. 1996b; Piatetsky-Shapiro & Frawley 1991). Model pengelasan boleh digunakan untuk mengelaskan sampel data yang baru sebaik mungkin bagi membantu proses memahami isi kandungan pangkalan data (Kamber et al. 1997) yang menyimpan berjuta-juta data. Selain daripada itu, model pengelasan juga dapat mewakili jumlah data yang besar yang tersimpan di dalam pangkalan data. Pengelasan boleh digunakan dalam banyak aplikasi seperti menentukan kelulusan kredit, pemasaran produk, diagnosis perubatan dan sebagainya.

Kaedah aruhan berasaskan atribut merupakan kaedah yang dibina bagi penemuan pengetahuan dalam pangkalan data hubungan. Kaedah ini bersepadu dengan paradigma pembelajaran mesin (Michalski 1983), terutamanya teknik

pembelajaran daripada contoh, operasi pangkalan data dan ekstraksi pengetahuan daripada pangkalan data. Walaupun mempunyai jumlah data yang banyak tetapi data tersebut tidak dapat menyediakan pengetahuan yang diperlukan bagi digunakan dalam proses membuat keputusan. Pada masa kini, dengan teknologi yang serba canggih dan maju, pengetahuan yang boleh mewakili data yang banyak serta dapat membantu proses membuat keputusan diperlukan. Adalah tidak mustahil untuk memperolehi jumlah petua yang banyak dengan menggunakan pelbagai teknik yang tidak kemas, berkemungkinan mengandungi petua-petua yang tidak bererti, bertindih antara satu sama lain ataupun petua yang berlebihan dalam model. Situasi ini akan menyebabkan proses membuat keputusan menjadi sukar dan rumit. Cara yang terbaik adalah menggunakan satu teknik yang hanya mengekstrak petua teritlak dan ringkas, yang boleh mewakili masalah sebenar tetapi masih mengekalkan kualiti pengetahuan dalam proses membuat keputusan yang baik. Konsep pengitlakan dan ringkasan adalah penting dalam memperolehi pengetahuan yang berguna dan bermanfaat dalam apa jua kerja yang berkaitan dengan pengurusan maklumat agar proses membuat keputusan dapat berjalan dengan lancar tanpa keraguan. Kertas kerja ini menerangkan mengenai pembinaan dua algoritma bagi melaksanakan kaedah aruhan berasaskan atribut dengan menggunakan pangkalan data hubungan untuk memperolehi petua yang bernilai dan berguna yang dapat mewakili jumlah data yang besar tersimpan dalam pangkalan data.

KAEDAH ARUHAN BERASASKAN ATRIBUT

Terdapat tiga primitif yang perlu ada dalam melaksanakan kaedah aruhan berasaskan atribut (Cai et al. 1990; Han et al. 1992; Han et al. 1993; Han et al. 1994a; Han et al. 1994b; Cheung et al. 2000). Pertama, pengelasan data berkaitan yang mana data pelajar siswazah UKM akan dikelaskan mengikut keputusan PNGK masing-masing. Kedua ialah pengetahuan asas iaitu maklumat mengenai domain yang hendak dilombong dan digunakan dalam proses penemuan pengetahuan yang biasanya diwakili dalam bentuk hierarki konsep (Han et al. 1992). Bagi kajian ini, hierarki konsep digambarkan seperti dalam Jadual 1. Ketiga pula ialah perwakilan keputusan pembelajaran iaitu hasil yang diperolehi daripada proses pengitlakan boleh digambarkan sebagai satu bentuk hubungan atau predikat kalkulus tertib.

Bagi melaksanakan kaedah aruhan berasaskan atribut, terdapat enam langkah asas yang perlu dilalui (Han et al. 1992), iaitu:

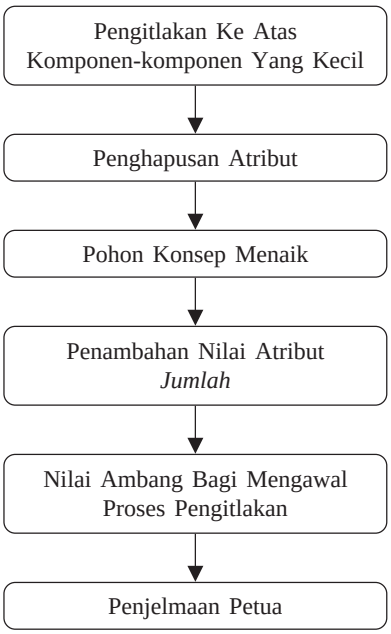
1. Pengitlakan ke atas setiap komponen-komponen yang kecil dalam hubungan data.
2. Penghapusan atribut iaitu atribut yang tidak mempunyai konsep tahap tinggi perlu dihapuskan walaupun atribut ini mempunyai jumlah nilai yang besar.

JADUAL 1. Jadual hieraki konsep bagi data pelajar

{Utara, Barat, Timur, Selatan, Tengah} ⊂ ANY(Kawasan)
{Perlis, Kedah, P.Pinang} ⊂ Utara
{Perak, Selangor} ⊂ Barat
{Kelantan, Terengganu, Pahang, Sabah, Sarawak} ⊂ Timur
{Melaka, N.Sembilan, Johor} ⊂ Selatan
{WPKL, WPPJ} ⊂ Tengah
{Padang Besar, Arau, Kangar} ⊂ Perlis
{Langkawi, Kubang Pasu, Padang Terap, Kota Setar, …}⊂ Kedah
{Barat Daya, Timur Laut, Butterworth, Kepala Batas, …}⊂ P.Pinang
{Senarai Daerah…} ⊂ Senarai Negeri
{Senarai Bandar…} ⊂ Senarai Daerah
{Sistem Maklumat Pengurusan, Sains Maklumat, Sains Komputer, Sains Industri, Kerja Kursus, Tesis) ⊂ ANY(Teknologi Maklumat)
{0.0-1.99} ⊂ Lemah
{2.0-2.99} ⊂ Baik
{3.0-3.49} ⊂ Sangat Baik
{3.5-4.0} ⊂ Cemerlang
{Lemah, Baik, Sangat Baik, Cemerlang} ⊂ ANY(PNGK)
{Lelaki, Perempuan} ⊂ ANY(Jantina)

3. Pengitlakan menggunakan pohon konsep menaik iaitu jika wujud konsep tahap tinggi dalam pohon konsep bagi sesuatu atribut, maka atribut tersebut boleh diitlakkan pada tahap tinggi berdasarkan pohon konsep menaik. Pengitlakan yang minimum seharusnya mengikut susunan menaik pohon pada satu tahap dan satu masa. Sebabnya, langkah ini selaras dengan petua pengitlakan pohon pengitlakan menaik dalam pembelajaran daripada contoh (Michalski 1983).
4. Penambahan nilai atribut *Jumlah* di mana atribut ini akan diwujudkan untuk mencatat jumlah rekod yang mempunyai nilai atribut yang sama dan disatukan. Atribut *Jumlah* ini sentiasa dibawa dalam setiap proses pengitlakan bagi merekodkan jumlah rekod yang disatukan.
5. Penetapan nilai ambang yang digunakan bagi mengawal proses pengitlakan. Jika jumlah rekod yang telah melalui proses pengitlakan lebih besar daripada nilai ambang yang telah ditetapkan, maka proses pengitlakan perlu dilaksanakan sekali lagi sehingga bilangan rekod menjadi lebih kecil atau sama dengan nilai ambang yang telah ditetapkan. Nilai ambang yang sesuai perlu ditetapkan agar proses pengitlakan tidak terlebih yang boleh menyebabkan kehilangan pengetahuan yang bernilai atau menyebabkan keputusan tidak teritlak sepenuhnya.
6. Penjelmaan petua iaitu satu rekod dalam hubungan akhir yang telah teritlak, dijelmakan kepada bentuk normal konjungsi manakala rekod berganda dijelmakan kepada bentuk normal disjungsi. Selain daripada itu rekod-rekod

yang teritlak boleh juga digambarkan dalam bentuk jadual hubungan. Rajah 1 menunjukkan langkah-langkah melaksanakan kaedah aruhan berasaskan atribut.



RAJAH 1. Langkah-langkah dalam melaksanakan kaedah aruhan berasaskan atribut

ALGORITMA DAN CARTA ALIR

Terdapat dua algoritma yang memainkan peranan penting dalam pembinaan model pengelasan iaitu:

1. Algoritma Kaedah Aruhan Berasaskan Atribut

Algoritma ini merupakan algoritma utama yang akan menciptakan satu objek kawalan pangkalan data yang dipanggil *DbController* yang akan berfungsi untuk menerima mesej (serta parameter, jika ada). Mesej-mesej yang diterima itu ialah mesej *SelectData*(parameter), mesej *InsertData*(parameter) dan mesej *UpdateData*(parameter). Objek ini juga akan menerima mesej *getConnection* bagi mendapatkan hubungan ke pangkalan data. Algoritma ini hanya perlu menghantar sebarang mesej kepada algoritma arahan bahasa piawai (*SQL*) melalui objek *DbController* yang diwujudkan. Algoritma 1 menunjukkan algoritma kaedah aruhan berasaskan atribut.

```
Input:   a) Satu pangkalan data hubungan
         b) Hieraki konsep
Output: Model pengelasan daripada pangkalan data.

Langkah 1: Kumpulkan semua data yang berkaitan.
Langkah 2: Sediakan pangkalan data beserta dengan jadual-jadual yang berisi data
           yang diperlukan.
Langkah 3: Semak atribut yang tidak mempunyai konsep tahap tinggi dalam jadual hieraki,
           jika ada atribut ini perlu dihapuskan sebelum melaksanakan
           pendekatan berasaskan atribut arahan.
Langkah 4: Laksanakan pendekatan berasaskan atribut aruhan.
Langkah 5: Ringkaskan hubungan yang telah diitlakan.

class mining{
void main(String[] args) {
Mula {pendekatan berasaskan atribut aruhan}

try
  {
        //mencipta satu objek baru daripada kelas DbController
        DbController db = new DbController();
        //menghantar mesej kepada objek DbController
        Connection connect = db.getConnection();
        (mula proses 1)
        //bina vektor baru
        Vector senarai=new Vector();
        //menghantar mesej memilih senarai data daripada pangkalan data dan mengira
        //bilangan rekod yang disatukan kepada objek DbController
        senarai = db. Select Data("namaJadual","namaMedan","namaMedan1");
        //dapatkan saiz vektor senarai
        int bilsize = senarai.size();
        //dapatkan bilangan rekod dengan bahagikan saiz vektor dengan bilangan medan
        int bilrekod = senarai.size()/bilangan medan;

for (bilangan rekod)
        {
                //dapatkan nilai medan dan tukar kepada string untuk disimpan dalam
                pangkalan //data.
                String namamedan = senarai.get(bil*2+0).toString();
                String namamedan1 = senarai.get(bil*2+1).toString();
                //hantar mesej kepada objek DbController untuk menyimpan data
                boolean data = b.InsertData("namaJadual","namaMedan","namaMedan1",
                nilaimedan, nilaimedan1);
                if(data1== true)
                System.out.println("DATA TELAH DISIMPAN");

        }
        Tandakan rekod yang bertindan dalam kelas atribut PNGK yang berbeza
                b.UpdateData("namaJadual","namaMedan",nilaiMedan,"namaMedan1",
                nilaiMedan1,"namaMedan2",nilaiMedan2)
        if bilangan rekod yang tidak bertindan > nilai ambang
                Ulangi proses 1
        else
                break;
}Tamat
 (tamat proses 1)
```

ALGORITMA 1. Algoritma kaedah aruhan berasaskan atribut

2. Algoritma Arahan Bahasa Pertanyaan Piawai

Algoritma ini mengandungi kelas *DbController* yang mengandungi segala perlakuan seperti membina hubungan ke pangkalan data, *loading driver*, membina pernyataan dan juga mengumpul semua mesej arahan bahasa pertanyaan piawai (*SQL*). Apabila kelas *DbController* menerima mesej daripada algoritma kaedah aruhan berasaskan atribut maka ia akan melaksanakan mesej tersebut dan akan mengembalikan output yang dikehendaki kepada algoritma kaedah aruhan berasaskan atribut. Antara arahan bahasa pertanyaan piawai ialah *SelectData*(parameter), *InsertData*(parameter) dan *UpdateData*(parameter). Rujuk Algoritma 2 iaitu algoritma arahan bahasa pertanyaan piawai.

```
Input:   NamaJadual, NamaMedan, NamaMedan1, nilaiMedan
Output:  Arahan Bahasa Pertanyaan Piawai (SQL)
class DbController() {
Mula
        Load the driver
        Bina hubungan dengan pangkalan data (getConnection)
        Bina pernyataan (create Statement)
        public Vector selectdata(namaJadual,namaMedan,namaMedan1) {
        //Pilih senarai data daripada pangkalan data berdasarkan parameter
        Select namaMedan,namaMedan1count namaMedan1 from
        namaJadual group by namaMedan,namaMedan1;
        try {
        while (bilangan rekod dalam pangkalan data) {
                String data1 = rs.getString(namaMedan);
                //Kumpulkan dalam satu vector A;
                vectorA.addElement(data1);
                String data2 = rs.getString(namaMedan1);
                vectorA.addElement(data2);}
                tutup hubungan ke pangkalan data;
        } catch()
        return vectorA ;
        public boolean insertdata(namaJadual,namaMedan,nilaiMedan)
        boolean insertData=false;
        try {//simpan data dalam jadual
        Insert into namaJadual(namaMedan) values (?);
        ps = connect.prepareStatement(query);
        ps.setString(1,nilaiMedan);
        ps.executeUpdate()
        commit;
        tutup hubungan ke pangkalan data;
        }catch()
        return insertData;
```

ALGORITMA 2. Algoritma arahan bahasa pertanyaan piawai (*SQL*)

Algoritma arahan bahasa pertanyaan piawai penting kerana segala urusan mengenai pangkalan data dikumpulkan dalam algoritma ini. Ketiadaan algoritma ini akan menyebabkan algoritma kaedah aruhan berasaskan atribut

menjadi perlahan dan juga panjang atur caranya. Secara asasnya, kedua-dua algoritma ini dilaksanakan dengan mencipta satu objek *DbController*. Kemudian, mesej dihantar kepada kelas *DbController* bagi mendapatkan hubungan ke pangkalan data. Setelah mesej diterima, kelas *DbController* akan membina hubungan dengan pangkalan data. Selepas itu, mesej *SelectData* pula akan dihantar kepada kelas *DbController* bagi mendapatkan senarai data serta bilangan rekod yang perlu disatukan jika mempunyai nilai atribut yang sama daripada pangkalan data. Senarai data akan dikembalikan dalam satu bentuk vektor. Saiz vektor dan bilangan rekod dalam vektor tersebut akan dikenal pasti dan berdasarkan bilangan rekod tersebut, senarai rekod akan ditukarkan daripada bentuk vektor kepada bentuk *string* supaya boleh disimpan dalam pangkalan data. Senarai rekod tersebut kemudiannya disimpan di dalam jadual pangkalan data. Rekod-rekod yang bertindan tetapi mempunyai kelas atribut PNGK yang berlainan akan ditandakan dengan simbol '*'. Akhir sekali, bilangan baris rekod yang tidak bertanda '*' perlu dikenal pasti. Sekiranya bilangan baris rekod lebih besar daripada nilai ambang, maka proses pengitlakan ini perlu diteruskan sehingga bilangan baris rekod lebih kecil atau sama dengan nilai ambang.

PENGUTIPAN DATA

Data mentah diperolehi daripada Fakulti Teknologi dan Sains Maklumat (FTSM), UKM. Sebanyak 204 data dan 25 atribut diperolehi. Setelah proses pembersihan data dilakukan, hanya 5 atribut digunakan untuk melaksanakan kaedah aruhan berasaskan atribut iaitu atribut *No_Matrik*, atribut *Kawasan*, atribut *Jabatan*, atribut *Jantina* dan atribut *PNGK*. Manakala, hanya 170 data diambil kira dalam kajian ini kerana selebihnya adalah data yang berkaitan dengan pelajar siswazah yang mengambil kursus tesis sepenuh masa. Tindakan ini dilakukan kerana tujuan kajian ini adalah untuk mengetahui prestasi pelajar siswazah berdasarkan pencapaian purata nilai kumulatif gred atau PNGK mengikut kawasan dan jabatan. Data pelajar siswazah tesis sepenuh masa tidak diambil kira kerana prestasi mereka tidak ditentukan berdasarkan PNGK, hanya keputusan lulus atau gagal sahaja yang menentu pencapaian mereka. Jadual 2 menunjukkan sebahagian daripada data pelajar yang digunakan dalam kajian ini, manakala Jadual 3 pula memaparkan penerangan mengenai setiap atribut yang terlibat dalam proses pengitlakan ini.

KEPUTUSAN DAN ANALISIS

Semasa uji kaji, data pelajar siswazah dibahagikan kepada 10 set (Set 1 hingga Set 10) mengikut kaedah *k-lipatan* pengesahan sahihan silang (Han & Kamber 2001). Setiap set data berbeza mengikut pecahan data latihan dan data ujian mengikut pecahan peratusan. Dalam setiap set data pula disediakan

JADUAL 2. Sebahagian daripada set data pelajar siswazah FTSM, UKM

Bil. Pelajar	Kawasan	PNGK	Jabatan
1	Timur	Sangat Baik	TP
2	Barat	Sangat Baik	TS
3	Utara	Cemerlang	TK
4	Selatan	Cemerlang	KK
5	Selatan	Sangat Baik	TK
..	..	..	..
170	Utara	Lemah	TP

JADUAL 3. Penerangan mengenai atribut

Atribut	Keterangan
Kawasan	Pelajar dibahagikan kepada 5 kawasan iaitu utara, barat, timur, selatan dan tengah. Pembahagian ini adalah berdasarkan alamat asal pelajar.
PNGK	Purata Nilai Gred Kumulatif atau PNGK ialah keputusan akademik pelajar yang dibahagikan kepada 4 kelompok iaitu Cemerlang (3.5-4), Sangat Baik (3.0-3.49), Lemah (2.0-2.9) dan Gagal (0.0-1.99).
Jabatan	Pelajar dibahagikan kepada beberapa jabatan iaitu Jabatan Pengurusan Sistem (TS), Jabatan Sains Maklumat (TP), Jabatan Sains Komputer (TK) dan Jabatan Kerja Kursus (KK).

10 sampel data (Sampel 1 hingga Sampel 10) bertujuan memastikan setiap data berpeluang untuk dilatih dan diuji. Ini bermakna, sejumlah 100 set data menjalani proses saling latih dan uji. Bagi memilih sampel data yang mempunyai ketepatan pengelasan maksimum daripada setiap set data, nilai ambang ketepatan pengelasan perlu ditentukan. Dalam uji kaji ini, nilai ambang ketepatan pengelasan ialah 70%. Jadual 4 menunjukkan beberapa contoh petua yang diperolehi hasil daripada perlombongan data menggunakan teknik aruhan berasaskan atribut. Dalam jadual tersebut, *d* merupakan nilai pemberat yang dihasilkan di mana petua akan diambil kira sekiranya nilai pemberat *d* didapati melebihi nilai ambang yang ditetapkan. Contohnya, daripada satu set data, sebanyak 25 daripada 30 petua yang dihasilkan mempunyai nilai pemberat melebihi nilai ambang. Oleh itu 25 petua ini akan digunakan sebagai pengetahuan untuk pengelasan kelak. Merujuk kepada Petua 3 dalam Jadual 4, didapati pelajar Jabatan KK dari kawasan barat agak lemah pencapaiannya tetapi dengan nilai pemberat yang kecil iaitu 14.28%. Oleh itu, petua ini tidak akan digunakan sebagai pengetahuan.

Hasil numerikal uji kaji yang diperolehi ditunjukkan dalam Jadual 5. *Ketepatan Maksimum* diperolehi daripada sampel yang memberikan ketepatan tertinggi dalam setiap set data, manakala *Purata Ketepatan* merupakan purata

JADUAL 4. Contoh petua dari model pengelasan yang dihasilkan

Petua 1:
$\forall$(x) (Cemerlang(x) $\rightarrow$ ((Kawasan(x) $\in$ Barat) $\wedge$ (Jabatan(x) $\in$ TS) [*d*:50%] $\vee$ ((Kawasan(x) $\in$ Barat) $\wedge$ (Jabatan(x) $\in$ TP) [*d*:50%] $\vee$ ((Kawasan(x) $\in$ Barat) $\wedge$ (Jabatan(x) $\in$ KK) [*d*:50%] $\vee$ ((Kawasan(x) $\in$ Barat) $\wedge$ (Jabatan(x) $\in$ TK) [*d*:50%]))
Petua 2:
$\forall$(x) (Lemah(x) $\rightarrow$ ((Kawasan(x) $\in$ Barat) $\wedge$ (Jabatan(x) $\in$ KK) [*d*:16.67%]
Petua 3:
$\forall$(x) (Lemah(x) $\rightarrow$ ((Kawasan(x) $\in$ Barat) $\wedge$ (Jabatan(x) $\in$ KK) [*d*:16.67%]$\vee$ ((Kawasan(x) $\in$ Timur) $\wedge$ (Jabatan(x) $\in$ KK) [*d*:25%]))

ketepatan semua 10 sampel dalam setiap set data. *Bilangan Petua Maksimum* ialah bilangan petua yang tertinggi di antara semua sampel manakala *Bilangan Petua Terpilih* pula ialah bilangan petua dari sampel yang memberi ketepatan tertinggi dalam setiap set data. *Purata Bilangan Petua* ialah purata bilangan petua yang dihasilkan dari semua sampel di setiap set data. Hasil keseluruhan uji kaji ke atas sampel tidak ditunjukkan untuk meringkaskan kertas ini. Daripada hasil proses pengitlakan bagi semua set data, hanya 5 set data yang mempunyai nilai ketepatan pengelasan yang lebih besar daripada nilai ambang yang telah ditetapkan iaitu Set 3, Set 4, Set 5, Set 6 dan Set 10. Bagi Set 10, walaupun mempunyai ketepatan pengelasan tertinggi di antara 5 set data tersebut iaitu 82.35%, tetapi kerana peratus bagi data latihan adalah lebih tinggi daripada peratus untuk data ujian, maka set data tersebut tidak boleh diambil kira kerana sudah semestinya banyak petua yang terhasil yang akan dapat memenuhi sebahagian data ujian. Keputusan yang memberangsangkan boleh dilihat dari Set 3 hingga Set 6 kerana mempunyai ketepatan pengelasan maksimum yang agak baik iaitu melebihi 70%. Walaupun mempunyai peratus data latihan yang sedikit, set-set ini mampu menghasilkan petua yang dapat memberi peratus ketepatan pengelasan yang baik.

Oleh itu, perhatian hanya akan diberikan kepada Set 3 hingga Set 6 kerana walaupun mempunyai peratus data latihan yang sedikit tetapi dapat memberi peratus ketepatan pengelasan yang baik. Ini menunjukkan bahawa jumlah petua yang dijana dan dipilih dapat mengelas dengan baik jumlah data ujian yang banyak. Bagi Set 4, walaupun peratus data latihan lebih rendah iaitu 40% daripada 60% data ujian, tetapi set ini boleh menghasilkan peratus ketepatan pengelasan maksimum, iaitu dari Sampel 7 yang mempunyai 78.43% ketepatan pengelasan. Begitu juga dengan Set 5 dan Set 6, di mana pembahagian data latihan dan data ujian adalah sama rata, iaitu 50% data

latihan dapat menghasilkan model pengelasan yang dapat mengelas dengan baik daripada sebahagian daripada 50% data ujian. Ini menunjukkan bahawa pembahagian data sama rata juga akan menghasilkan model pengelasan yang akan dapat mengelas data ujian dengan baik.

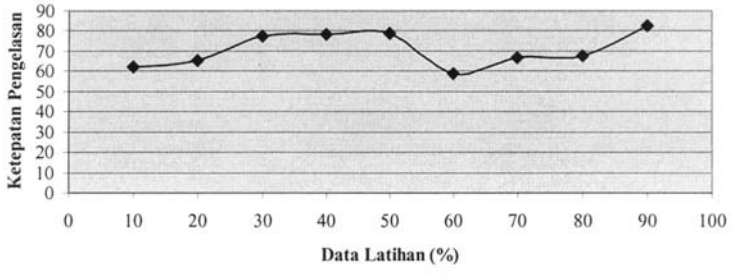
JADUAL 5. Hasil proses pengitlakan bagi semua set data

Set	Pecahan Peratus Latih: Ujian	Ketepatan Maksimum (%)	Purata Ketepatan (%)	Bilangan Petua Maksimum	Bilangan Petua Terpilih	Purata Bilangan Petua
1	10:90	62.09	49.76	16	16	15
2	20:80	65.44	53.97	22	18	23
3	30:70	77.31	60.34	31	25	31
4	40:60	78.43	61.86	32	30	32
5	50:50	78.82	64.47	36	24	34
6	50:50	78.82	59.88	34	18	35
7	60:40	58.82	53.23	37	20	35
8	70:30	66.67	57.06	40	23	36
9	80:20	67.65	63.53	37	23	39
10	90:10	82.35	61.18	40	23	40

Hasil uji kaji juga menunjukkan peratus ketepatan pengelasan maksimum bagi dua set data ini ialah 78.82% diperolehi dari Sampel 3 bagi Set 5 dan Sampel 10 bagi Set 6. Manakala Set 3, walaupun peratus data latihan adalah lebih rendah iaitu hanya 30% berbanding peratus data ujian yang lebih tinggi iaitu 70%, tetapi set ini mempunyai jumlah petua yang boleh mengelas jumlah data ujian yang banyak dengan baik. Sampel data bagi set 3 yang menghasilkan peratus ketepatan pengelasan maksimum ialah dari Sampel 9 di mana nilai ketepatannya ialah 77.31%. Set data yang mempunyai peratus data latihan dan data ujian yang sama rata tetapi mempunyai peratus ketepatan pengelasan yang tertinggi sekali ialah Set 5 dan Set 6. Oleh itu dapat dirumuskan bahawa pembahagian sama rata antara peratus data latihan dan data ujian juga akan dapat menghasilkan jumlah petua yang boleh mengelas dengan baik. Jumlah petua yang paling sedikit yang terpilih adalah bagi Set 6. Ini menguatkan lagi rumusan bahawa pembahagian sama rata peratus data latihan dan data ujian akan dapat menghasilkan jumlah petua yang sedikit tetapi dapat mengelas jumlah data ujian yang banyak dengan baik.

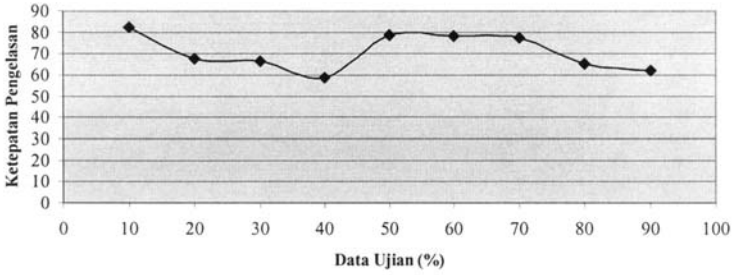
Rajah 1 menunjukkan hubungan peratus data latihan dengan ketepatan pengelasan. Didapati ketepatan pengelasan semakin meningkat apabila peratus data latihan semakin bertambah iaitu di antara 30% hingga 50%. Nilai ketepatan pengelasan menurun semasa peratus data latihan di antara 10% hingga 30%. Ini mungkin kerana petua yang dihasilkan adalah sedikit, menyebabkan peratus pengelasan menurun. Apabila peratus data latihan

bertambah di antara 50% hingga 80%, petua yang dihasilkan adalah banyak tetapi berdasarkan kepada Rajah 1, didapati peratus pengelasan menurun. Ini mungkin disebabkan oleh petua yang dihasilkan banyak yang bertindih ataupun ada petua yang tidak penting. Oleh itu, dapat dirumuskan bahawa tidak semestinya pertambahan nilai peratus data latihan akan dapat menghasilkan peratus ketepatan pengelasan yang maksimum kerana ada kemungkinan petua-petua yang dihasilkan bertindih atau tidak berkaitan.



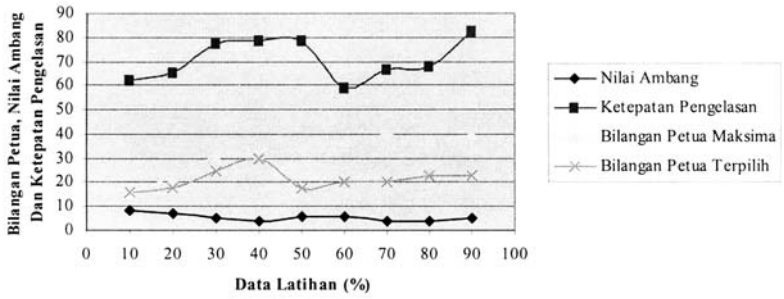
RAJAH 1. Hubungan peratus data latihan dengan ketepatan pengelasan

Rajah 2 pula menunjukkan bahawa nilai peratus ketepatan pengelasan semakin menurun apabila nilai peratus data ujian semakin bertambah, iaitu ketika peratus data ujian bernilai di antara 10% hingga 40% dan di antara 70% hingga 90%. Tetapi, pada ketika peratus data ujian bernilai di antara 50% hingga 70%, peratus ketepatan pengelasan mencapai tahap maksimum. Ini menunjukkan bahawa peratus data ujian juga memainkan peranan penting dalam menghasilkan peratus ketepatan pengelasan maksimum. Oleh itu, dapat disimpulkan bahawa pertambahan peratus data ujian boleh mempengaruhi nilai peratus ketepatan pengelasan sama ada tinggi ataupun rendah.



RAJAH 2. Hubungan peratus data ujian dengan ketepatan pengelasan

Bagi melaksanakan proses pengitlakan ini, nilai ambang sentiasa berubah-ubah sehinggalah nilai yang sesuai diperolehi pada peringkat proses pengitlakan mencecah tahap yang terakhir ataupun apabila tiada lagi rekod yang mempunyai nilai atribut yang sama yang perlu disatukan. Nilai ambang bagi pengujian kes ini merupakan jumlah rekod yang mempunyai nilai atribut yang tidak sama antara satu kelas dengan kelas-kelas yang lain, ataupun bermaksud jumlah rekod yang mempunyai nilai atribut yang unik. Penentuan nilai ambang ini penting kerana nilai ambang jika ditetapkan terlalu kecil, boleh menyebabkan banyak rekod yang teritlak, dan kemungkinan kehilangan maklumat yang bernilai dan berguna. Manakala, nilai ambang yang terlalu besar pula boleh menyebabkan banyak bilangan rekod yang tidak teritlak sepenuhnya dan petua yang terhasil ini mungkin tidak mampu mengelas dengan baik. Jika Rajah 3 diperhatikan, didapati peratus data latihan boleh mempengaruhi nilai ambang yang diperlukan bagi menghasilkan model pengelasan.



RAJAH 3. Hubungan peratus data latihan ke atas nilai ambang, bilangan petua maksimum dan terpilih serta ketepatan pengelasan

Merujuk kepada Rajah 3 dan Jadual 6, didapati nilai ambang semakin berkurangan dengan pertambahan peratus data latihan. Jika peratus data latihannya dalam lingkungan 10% hingga 20%, nilai ambangnya ialah di antara 7 hingga 8. Manakala bagi Set 3 hingga ke Set 9, didapati nilai ambang adalah di antara 4 hingga 6, dengan peratusan data latihan semakin meningkat. Kesimpulannya, jika peratus data latihan melebihi 20% daripada jumlah keseluruhan set data, nilai ambang boleh disetkan di antara 4 hingga 6.

Berdasarkan rumusan di atas, nilai ambang boleh dianggarkan berdasarkan peratus data latihan. Jika diperhatikan nilai ambang bagi Set 3, Set 4 dan Set 5 iaitu set-set yang mempunyai sampel yang menghasilkan ketepatan pengelasan maksimum, didapati set-set ini juga mempunyai nilai ambang di

JADUAL 6. Nilai ambang bagi semua set data

Set	Ketepatan Maksimum (%)	Bilangan Petua Maksimum	Bilangan Petua Terpilih	Nilai Ambang
1	62.09	16	16	8
2	65.44	22	18	7
3	77.31	31	25	5
4	78.43	32	30	4
5	78.82	34	18	6
6	58.82	37	20	6
7	66.67	40	23	4
8	67.65	37	23	4
9	82.35	40	23	5

antara 4 hingga 6. Kesimpulannya, peratus data latihan yang rendah akan mempunyai nilai ambang yang agak tinggi iaitu di antara 7 hingga 8, manakala peratus data latihan yang tinggi akan mempunyai nilai ambang yang agak rendah iaitu di antara 4 hingga 6. Berdasarkan daripada kajian ini, dapat dirumuskan bahawa nilai ambang yang bersesuaian dengan proses pengitlakan ini ialah di antara 4 hingga 6.

Peratus data latihan juga boleh mempengaruhi bilangan petua yang terhasil dan juga dalam meningkatkan peratus ketepatan pengelasan. Pertambahan peratus data latihan boleh menambahkan bilangan petua yang terhasil. Berdasarkan daripada kajian yang telah dilakukan, didapati peratus data latihan boleh mempengaruhi nilai ambang, bilangan petua yang terhasil serta dalam meningkatkan peratus ketepatan pengelasan. Kesimpulannya, peratus data latihan yang tinggi mempunyai nilai ambang yang rendah serta boleh mempengaruhi dalam meningkatkan bilangan petua yang terhasil dan yang terpilih dan juga dalam meningkatkan peratus ketepatan pengelasan. Manakala, peratus data latihan yang rendah mempunyai nilai ambang yang tinggi, dan boleh mempengaruhi dalam mengurangkan bilangan petua yang terhasil dan terpilih, serta peratus ketepatan pengelasan.

KESIMPULAN

Kertas kerja ini telah membuktikan bahawa kaedah aruhan berasaskan atribut boleh digunakan untuk menghasilkan petua yang bernilai serta boleh mewakili jumlah data yang besar yang disimpan dalam pangkalan data hubungan. Kajian ini juga menunjukkan bahawa peratus data latihan yang sedikit juga boleh menghasilkan petua yang bernilai dan berguna, serta dapat mengelas dengan baik jumlah maklumat yang disimpan dalam pangkalan data hubungan. Penentuan nilai ambang juga dipengaruhi oleh peratus data latihan iaitu satu

nilai yang digunakan untuk mengawal proses pengitlakan, dan sudah tentunya peratus data latihan memainkan peranan penting dalam menghasilkan bilangan petua sama ada sedikit atau banyak.

PENGHARGAAN

Penulis ingin memberi penghargaan kepada MIMOS Berhad kerana telah menyokong dan membiayai penyelidikan ini.

RUJUKAN

Cai, Y., Cercone, N. & Han, J. 1990. An attribute-oriented approach for learning classification rules from relational databases. *Proceedings of the 6th International Conference on Data Engineering*, 5-9 Februari. Los Angeles, 281-288.

Cai, Y., Cercone, N. & Han, J. 1991. Attribute-oriented induction in relational databases. Dlm. Piatetsky-Shapiro, G. & Frawley, W. J. (pnyt.). *Knowledge discovery in databases*, hlm. 213-228. Menlo Park, CA: AAAI/MIT Press.

Chen, M. S., Han, J. & Yu, P. S. 1996. Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering* 8(6): 866-883.

Cheung, D. W., Hwang, H. Y., Fu, A. W. & Han, J. 2000. Efficient rule-based attribute-oriented induction for data mining. *Journal of Intelligent Information Systems* 15(20): 175-200.

Fayyad, U. M., Piatetsky-Shapiro, G. & Symth, P. 1996a. Knowledge discovery and data mining: towards a unifying framework. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)*, 2-4 August. Portland, Oregon, USA, 82-88.

Fayyad, U. M., Piatetsky-Shapiro, G., Symth, P. & Uthurusamy, R. 1996b. From data mining to knowledge discovery: an overview. Dlm. Fayyad, U., Piatetsky-Shapiro, G., Symth, P. & Uthurusamy, R. (pnyt.). *Advances in knowledge discovery and data mining*, hlm. 1-35. Menlo Park, CA: AAAI/MIT Press.

Fu, Y. 1996. Discovery of multiple-level rules from large databases. Ph.D Thesis. Simon Fraser University, Barnaby, Canada.

Han, J., Cai, Y. & Cercone, N. 1992. Knowledge discovery in databases: an attribute-oriented approach. *Proceedings of the 18th International Conference on Very Large Data Bases (VLDB'92)*, 23-27 Ogos. Vancouver, Canada, 547-559.

Han, J., Cai, Y. & Cercone, N. 1993. Data-driven discovery of quantitative rules in relational database. *IEEE Transactions on Knowledge and Data Engineering* 5(1): 29-40.

Han, J., Cai, Y., Cercone, N. & Huang, Y. 1994a. Discovery of data evolution regularities in large databases. *Journal of Computer and Software Engineering* 3(1): 41-69.

Han, J., Fu, Y. & Ng, R. 1994b. Cooperative query answering using multiple layered databases. *Proceedings of the 2nd International Conference on Cooperative Information Systems (CoopIS'94)*, 17-20 Mei. Toronto, Canada, 47-58.

Han, J. & Kamber, M. 2001. *Data mining: concepts and techniques*. San Francisco, CA: Morgan Kaufman Publisher.

Kamber, M., Winstone, L., Gong, W., Cheng, S. & Han, J. 1997. Generalization and decision tree induction: efficient classification in data mining. *Proceedings of the 7th International Workshop on Research Issues on Data Engineering (RIDE'97)*, 7-8 April. Birmingham, UK, 111-120.

Michalski, R. S. 1983. A theory and methodology of inductive learning. Dlm. Michalski, R. S., Carbonell, J. G. & Mitchell, T. M. (pnyt.). *Machine learning: an artificial learning intelligence approach*, hlm. 83-134. Los Altos, CA: Morgan Kaufmann Publisher.

Piatetsky-Shapiro, G. & Frawley, W. J. 1991. *Knowledge discovery in databases*. Menlo Park, CA: AAAI/MIT Press.

Silberschatz, A., Stonebraker, M. & Ullman, J. D. 1991. Database system: achievements and opportunities. *Communications of the ACM* 33: 94-109.

Silberschatz, A., Stonebraker, M. & Ullman, J. D. 1996. Database research: achievements and opportunities into the 21st Century. *SIGMOD Record* 25: 52-63.

Turban, E. & Aronson, J. E. 2001. *Decision support systems and intelligent systems*. Ed. ke-6. New Jersey: Prentice Hall.

Siti Rohaidah Ahmad & Azuraliza Abu Bakar
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia
43600 UKM Bangi
Selangor Darul Ehsan
e-mail: aab@ftsm.ukm.my