APLIKASI TEKNIK PENDISKRITAN DALAM PERLOMBONGAN DATA

Nor Liyana Mohd Shuib
Faculty of Computer Science & Information Technology
University of Malaya, 50603, Kuala Lumpur, Malaysia.
Email: liyanashuib@gmail.com

Mohammad Hafiz Ismail
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, 02600, Arau, Perlis, Malaysia.
Email: MohammadHafiz@perlis.uitm.edu.my

Abstrak
Pendiskritan data merupakan kaedah pra-pemprosesan yang penting dalam membangunkan model pengelasan. Teknik pendiskritan data digunakan untuk menukarkan atribut selanjar kepada atribut diskrit. Ia sangat penting dalam membangunkan model berasaskan petua seperti pepohon keputusan dan set kasar. Penggunaan teknik pendiskritan dapat meningkatkan ketepatan pengelas dan menjadikan pembelajaran lebih tepat dan laju. Objektif kajian ini ialah untuk mengaplikasikan teknik pendiskritan data yang terpilih ke atas empat set data daripada UCI Machine Learning dan membuat perbandingan prestasi berdasarkan ketepatan pengelas, bilangan petua dan panjang petua. Teknik pendiskritan yang digunakan ialah teknik Taakulan Boolean, Equal Frequency Binning dan Entropi. Setiap teknik ini diaplikasikan ke atas empat set data dari domain yang berbeza untuk mendapatkan satu teknik yang terbaik. Set data tersebut ialah Iris, Glass, Pima dan Wine. Model pengelasan perlombongan data dibangunkan menggunakan kaedah pengelas set kasar melalui beberapa proses seperti pra-pemprosesan data, pembahagian set data latihan dan ujian, perlombongan data, pengujian dan perbandingan. Satu analisis perbandingan ke atas teknik pendiskritan yang digunakan dihasilkan. Hasil analisis mendapati penggunaan teknik Taakulan Boolean menggeneralisasikan purata ketepatan yang tertinggi jika dibandingkan dengan dua teknik yang lain.
Kata Kunci: *Pendiskritan Data, Pra-pemprosesan Data, Perlombongan Data*

1. Pengenalan

Salah satu teknik pra-pemprosesan yang sangat penting ialah teknik pendiskritan data. Set data dunia sebenar selalunya mengandungi atribut selanjar. Penggunaan atribut selanjar melibatkan saiz storan yang lebih besar, sukar difahami dan menghasilkan petua yang lebih panjang. Walau bagaimanapun, kebanyakan tugas pengelasan dunia sebenar yang wujud tidak dapat mengaplikasikan atribut selanjar selagi atribut selanjar ini tidak didiskritkan terlebih dahulu.

Atribut diskrit mempunyai peranan yang penting dalam perlombongan data. Atribut diskrit melibatkan selang nombor yang lebih ringkas untuk perwakilan, mudah digunakan dan mudah difahami kerana ia sangat rapat dengan perwakilan tahap pengetahuan berbanding dengan atribut selanjar. Petua daripada atribut diskrit selalunya lebih pendek dan mudah difahami. Penggunaan atribut diskrit dapat meningkatkan ketepatan ramalan dan menjadikan pembelajaran lebih tepat dan laju (Dougherty et al., 1995). Oleh itu, teknik pendiskritan data diperlukan untuk menukar atribut selanjar kepada atribut diskrit.

Teknik pendiskritan data boleh ditakrifkan sebagai satu proses mengkuantitikan atribut selanjar (Liu et al., 2002). Terdapat banyak teknik pendiskritan data yang wujud hasil daripada kajian yang meluas dalam bidang pendiskritan sejak kebelakangan ini (Nor Liyana Mohd Shuib et al., 2009). Pada mulanya, teknik-teknik pendiskritan data yang mudah digunakan seperti teknik binning iaitu *equal width* (EW) dan *equal-frequency* (EF). Sejajar dengan keperluan untuk mendapatkan pengelasan yang lebih tepat dan berkesan, teknologi untuk pendiskritan berkembang maju dengan pantas. Banyak teknik pendiskritan data telah dicadangkan dan diuji

bagi menunjukkan bahawa pendiskritan mempunyai potensi untuk mengurangkan jumlah data sambil mengekalkan atau meningkatkan ketepatan ramalan.

Disebabkan kepelbagaian teknik pendiskritan data, pelombong data menghadapi masalah untuk memilih teknik pendiskritan yang bersesuaian dengan permasalahan. Selalunya, pelombong data akan menguji beberapa teknik pendiskritan untuk menghasilkan data yang paling berkualiti. Objektif utama kajian ini ialah mengkaji teknik pendiskritan dalam perlombongan data. Berdasarkan objektif utama, objektif kajian yang lebih spesifik ialah untuk membuat perbandingan prestasi di antara tiga teknik pendiskritan data dalam model pengelasan berdasarkan ketepatan pengelas, bilangan petua dan panjang petua dan mengaplikasikan teknik pendiskritan data yang terpilih ke atas empat set data yang berbeza.

Dalam membuat perbandingan prestasi, tiga teknik pendiskritan data diaplikasikan ke atas empat set data daripada UCI *Machine Learning*. Setiap set data menjalani pra-pemprosesan data, pendiskritan data (menggunakan tiga teknik yang berbeza), pembahagian set data latihan dan ujian, perlombongan petua, pengujian dan perbandingan model dan petua yang dihasilkan. Tiga teknik yang digunakan sebagai teknik perbandingan ialah *Equal Frequency Binning* (EF), Taakulan Boolean (BR) dan teknik Entropi (Ent-MDLP).

Kertas ini terdiri daripada 6 bahagian. Bahagian pertama ialah kajian literatur mengenai teknik pendiskritan data. Seterusnya, bahagian kedua menerangkan pembangunan model perlombongan data yang digunakan untuk menjalankan analisis perbandingan. Kemudian, bahagian ketiga memaparkan hasil keputusan dan perbincangan. Bahagian terakhir pula membincangkan kesimpulan, sumbangan dan cadangan di masa hadapan.

2. Kajian Literatur

Pra-pemprosesan terdiri daripada empat kategori iaitu pembersihan data, integrasi data, transformasi data dan pengurangan data. Pembersihan data digunakan untuk menyingkirkan atau membetulkan data dengan mengisi nilai yang hilang, melicinkan data hingar, mengenal pasti atau membuang data asing dan menyelesaikan data yang tidak konsisten (Han & Kamber, 2001). Integrasi data pula diperlukan bagi menyelaraskan penggabungan data dari pelbagai sumber. Transformasi data ialah penukaran data ke bentuk atau format yang sesuai untuk perlombongan data. Pengurangan data pula diperlukan untuk mengurangkan saiz data yang besar kepada saiz yang lebih kecil tetapi tetap menghasilkan analisis yang sama. Pengurangan data penting kerana memproses data yang bersaiz kecil adalah lebih mudah daripada memproses data yang bersaiz besar (Panda, 2005).

Satu daripada bentuk kaedah pengurangan data ialah pendiskritan data. Pendiskritan data berfungsi untuk menukarkan atribut selanjar kepada atribut diskrit (Goharian & Grossman, 2003). Oleh kerana kebanyakan algoritma perlombongan hanya menerima atribut diskrit sahaja maka atribut selanjar perlu ditukarkan ke atribut diskrit terlebih dahulu. Kelebihan pendiskritan data ialah dapat menjadikan pembelajaran lebih tepat dan laju serta dapat menghasilkan keputusan (pemberat, petua) yang lebih padat, pendek dan tepat berbanding dengan menggunakan data selanjar (Dougherty et al., 1995).

2.1 Proses Pendiskritan Data

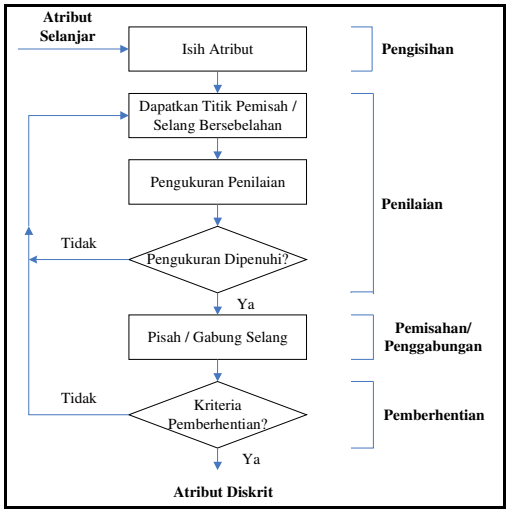
Proses pendiskritan yang dihuraikan ini merupakan proses pendiskritan asas iaitu hanya melibatkan pendiskritan secara univariat sahaja. Pendiskritan univariat hanya mempertimbangkan satu atribut selanjar sahaja pada satu masa manakala pendiskritan secara pelbagai variat (*multivariate)* pula mempertimbangkan beberapa atribut secara serentak.

Kebanyakan teknik pendiskritan menggunakan proses pendiskritan ini sebagai asas dan beberapa pembaikan dilakukan mengikut keperluan teknik tersebut. Menurut Liu et al. (2002), proses

pendiskritan terdiri daripada empat (4) langkah asas iaitu pengisihan, penilaian, pemisahan/penggabungan dan akhir sekali ialah pemberhentian. Proses pendiskritan digambarkan dalam Rajah 1.

Proses pendiskritan dimulakan dengan langkah pengisihan. Nilai selanjar sesuatu atribut diisih sama ada dalam bentuk susunan menaik atau menurun. Pemilihan algoritma pengisihan yang sesuai adalah penting untuk meningkatkan kelajuan proses pendiskritan contohnya seperti *Quick Set* yang merupakan algoritma penyusunan yang berkesan. Cara lain untuk meningkatkan keberkesanan adalah dengan mengelakkan pengisihan nilai atribut yang berulang-ulang.

Selepas pengisihan, langkah seterusnya ialah penilaian. Langkah penilaian mencari titik pemisah terbaik untuk pemisahan selang atau mencari pasangan selang bersebelahan yang terbaik untuk penggabungan. Salah satu fungsi penilaian adalah untuk menentukan hubungan pemisahan dan penggabungan selang dengan label kelas. Terdapat beberapa fungsi penilaian seperti pengukuran entropi dan pengukuran statistik.



Rajah 1: Proses pendiskritan data

Langkah ketiga ialah pemisahan/penggabungan. Dalam pemisahan atau juga dikenali sebagai pendekatan atas-bawah, selang dipisahkan manakala dalam penggabungan atau pendekatan bawah-atas, selang digabung. Untuk proses pemisahan, titik pemisah yang terbaik dipilih untuk memisahkan selang nilai selanjar kepada dua bahagian. Pendiskritan berterusan (bertambah satu demi satu) sehingga kriteria pemberhentian ditemui. Proses penggabungan juga melaksanakan teknik yang sama di mana selang bersebelahan dinilai untuk mencari pasangan selang terbaik untuk digabungkan dalam setiap iterasi. Pendiskritan diteruskan dengan menggabungkan selang sehingga kriteria pemberhentian ditemui. Langkah terakhir iaitu langkah pemberhentian ditentukan oleh kriteria pemberhentian. Kriteria pemberhentian menentukan syarat untuk memberhentikan proses pendiskritan. Kriteria pemberhentian yang mudah ialah seperti menetapkan bilangan selang pada awal proses atau yang lebih kompleks seperti menilai fungsi.

2.2 Teknik Pendiskritan Data Untuk Pembangunan Model

Teknik pendiskritan data yang dipilih ialah *Equal Frequency Binning* (EF), Taakulan Boolean (BR) (Nguyen & Skowron, 1997; Pawlak & Skowron, 2007) dan teknik berasaskan Entropi iaitu Ent-MDLP (Fayyad & Irani, 1993). Teknik EF dan Ent-MDLP dipilih kerana kedua-duanya banyak digunakan sebagai perbandingan di dalam kajian pendiskritan (Dougherty et al., 1995; Gama & Pinto, 2006). Teknik EF merupakan teknik pendiskritan tidak diselia yang paling asas

manakala teknik Ent-MDLP merupakan antara teknik pendiskritan yang terbaik (Cerquides & López de Màntaras, 1997; Liu & Wang, 2005; Tay & Shen, 2002). Gama & Pinto (2006) membandingkan teknik EF, EW dan MDLP dalam kajiannya. Hasil kajian mendapati Ent-MDLP menghasilkan keputusan yang terbaik.

Teknik BR dipilih kerana teknik ini sesuai untuk pengelas set kasar dan merupakan pendekatan yang baik untuk kajian yang melibatkan pengelasan dan pengecaman (Pawlak & Skowron, 2007). Ini adalah kerana teknik ini menggunakan pengukuran set kasar sebagai asasnya. Teknik ini juga belum digunakan secara meluas oleh para penyelidik sebagai salah satu teknik bagi perbandingan teknik pendiskritan. Ini merupakan salah satu sebab perbandingan prestasi dilaksanakan menggunakan teknik BR, EF dan Ent-MDLP.

2.2.1 Teknik Pendiskritan EF

Teknik EF merupakan salah satu daripada teknik binning (Catlett, 1991; Ching et al., 1995; Dougherty et al., 1995; Kerber, 1992; Kotsiantis et al., 2006; Li & Cercone, 2005; Ventura & Martinez, 1995). Teknik ini mendiskritkan atribut selanjar dengan menentukan bilangan bin menggunakan frekuensi. Bilangan selang, *k,* digunakan untuk menentukan bilangan bin. *k* ialah parameter yang ditakrifkan pengguna.

Teknik EF dimulakan dengan mengisih nilai daripada atribut selanjar. Teknik ini kemudiannya membahagikan nilai yang diisih kepada *k* selang supaya setiap selang mengandungi bilangan objek yang sama (diberi *n* objek). Setiap selang mengandungi *n/k* objek. Nilai yang sama mestilah diletakkan dalam selang yang sama. Label selang yang terhasil akan digunakan sebagai nilai baru. Kriteria pemberhentian tidak diperlukan dalam teknik ini kerana bilangan bin adalah tetap. Sebagai contoh, terdapat atribut selanjar iaitu atribut umur yang mempunyai nilai 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34 (selepas diisih). Diberi nilai *k* ialah tiga (3). Maka dengan menggunakan EF, atribut selanjar dibahagikan kepada tiga selang atau bin.

2.3.2 Teknik Pendiskritan Taakulan Boolean (BR)

Teknik yang seterusnya ialah teknik pendiskritan Taakulan Boolean (BR) yang dicadangkan oleh Nguyen dan Skowron (1997). Teknik ini juga digunakan oleh Pawlak & Skowron (2007) dalam teori set kasar. BR dibangunkan berasaskan teori set kasar dan Taakulan Boolean. Teknik ini merupakan teknik pendiskritan diselia yang mempertimbangkan kesemua atribut secara serentak dan menghasilkan titik pemisah yang lebih sedikit.

Andaikan L ialah satu set pasangan objek dengan kelas keputusan yang berbeza, contohnya, $L = \{\langle x, y\rangle \in U \times U \mid d(x) \neq d(y)\}$ dan $v_1^a < ... < v_j^a ... < v_{m_a}^a$ ialah nilai terisih atribut a seperti sebelumnya. Teknik ini memilih subset semi-minimal P daripada set kesemua kemungkinan titik pemisah (ke atas kesemua atribut) menggunakan $C = \left\{ \left\langle a, \frac{v_j^a + v_{j+1}^a}{2} \right\rangle \mid a \in A \;\; dan \;\; j = 1,...,(m_{a-1}) \right\}$ iaitu bagi setiap pasangan $\langle x, y\rangle \in L$ terdapat pasangan $\langle a, c_i^a \rangle \in P$ di mana c_i^a membahagikan x dan y (contoh: $(a(x) < c_i^a < a(y))$ *atau* $(a(y) < c_i^a < a(x))$. P dijumpai menggunakan algoritma tamak yang memilih titik pemisah secara iteratif.

Bagi setiap iterasi, titik pemisah yang mempunyai perbezaan paling tinggi di antara bilangan pasangan objek dalam L dipilih dan ditambah ke dalam P. Pasangan yang dipisahkan menggunakan titik pemisah ini dipindahkan daripada *L.* Teknik ini berhenti apabila *L* kosong.

2.3.3 Teknik Pendiskritan Ent-MDLP

Teknik pendiskritan terakhir yang dipilih ialah Ent-MDLP. Ent-MDLP merupakan teknik pendiskritan berasaskan entropi yang dicadangkan oleh Fayyad & Irani (1993). Ent-MDLP menggunakan *entropy minimization heuristic* (EMH) untuk mendiskritkan nilai selanjar kepada beberapa selang. Teknik ini juga menggunakan kriteria *minimum description length* (Rissanen, 1978) untuk mengawal bilangan selang yang dihasilkan ke atas ruang selanjar. Teknik yang diselia ini menggunakan maklumat kelas entropi untuk memilih titik pemisah.
Jika diberi set objek *S*, atribut *A* dan titik pemisah *T*, *T* membahagikan set data *S* kepada subset S_1 dan S_2. Contohnya terdapat *k* kelas yang diwakilkan sebagai $C_1, C_2,...,C_k$. $P(C_i, S_j)$ pula menjadi pembahagian bagi set data dalam S_j yang ada kelas *i*. Kelas entropi bagi subset S_j ditakrifkan sebagai:

$$Ent(S) = -\sum_{i=1}^{k} P(C_i, S)\log(P(C_i, S))$$

dan maklumat entropi kelas $E(A, T; S)$ ialah:

$$E(A,T;S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

di mana:
$Ent(S_j)$ = entropi kelas untuk subset S_j
S_j = subset daripada S
C_i = kelas i
$P(C_i, S_j)$ = Pembahagian objek daripada S_j yang berada dalam kelas C_i
$E(A, T_A; S)$ = maklumat pembahagian entropi kelas oleh titik pemisah T_A dalam A
A = Atribut selanjar di mana pembahagian dijalankan
$|S_k|$ = bilangan objek dalam S_k.
Titik pemisah terbaik adalah titik yang memberi maklumat kelas entropi yang paling minima di antara semua calon titik pemisah. Pendiskritan binari hanya memisahkan selang kepada dua, Fayyad & Irani (1993) Ent-MDLP menjana algoritma untuk menghasilkan beberapa selang secara serentak dengan menggunakan *EMH* secara rekursif sehingga kriteria pemberhentian ditemui. Kriteria pemberhentian ini dikenali sebagai *minimum description length principle* (MDLP). MDLP diterangkan dalam rumus berikut:

$$Gain(A,T;S) > \frac{\log_2(N-1)}{N} + \frac{\Delta(A,T;S)}{N}$$

di mana

$$\begin{aligned} Gain(A,T;S) &= Ent(S) - E(A,T;S) \\ &= Ent(S) - \frac{|S_1|}{N} Ent(S_1) - \frac{|S_2|}{N} Ent(S_2) \end{aligned}$$

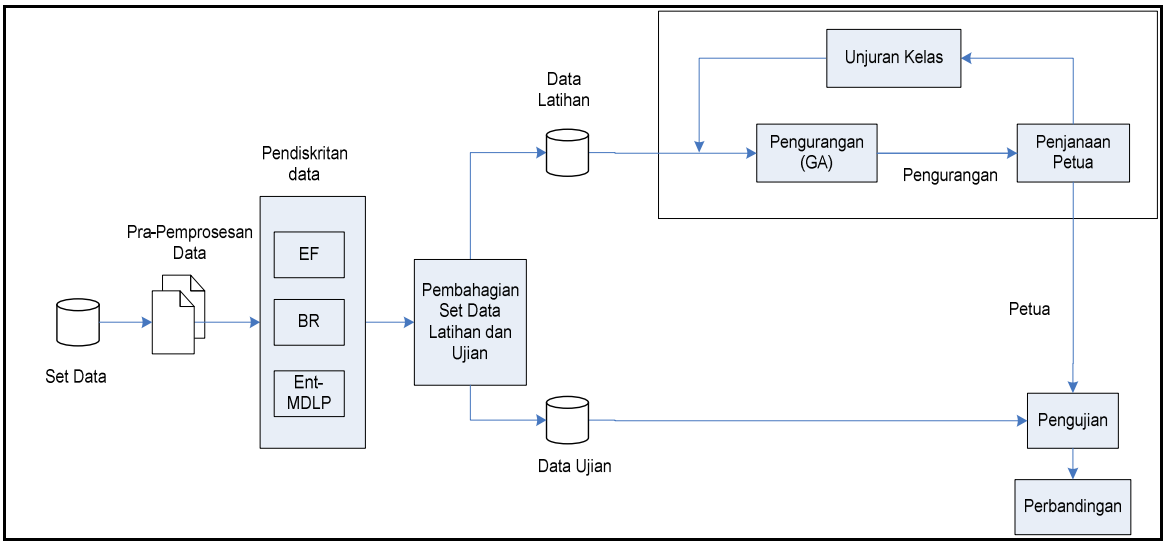
$$\Delta(A,T;S) = \log_2(3^k - 2) - [kEnt(S) - k_1 Ent(S_1) - k_2 Ent(S_2)]$$

N ialah bilangan objek dalam set *S*.

3. Pembangunan Model Pengelasan

Untuk membangunkan model pengelasan, pengelas set kasar telah dipilih sebagai teknik perlombongan data. Pengelas set kasar (Pawlak, 1982) dipilih kerana ia memerlukan atribut diskrit untuk melombong petua. Ini bersesuaian dengan fokus utama kajian ini iaitu untuk

membandingkan teknik pendiskritan data. Rajah 2 di bawah menunjukkan metodologi pembangunan model yang digunakan ke atas ketiga-tiga teknik pendiskritan yang dipilih:



Rajah 2: Pembangunan model pengelasan

Pembangunan model terdiri daripada penyediaan data (pra-pemprosesan data dan pembahagian set data latihan dan ujian), perlombongan data, pengujian dan perbandingan model dan petua yang dihasilkan.

3.1 Penyediaan Data

Penyediaan set data merupakan satu langkah yang penting dalam proses perlombongan data. Matlamat utama penyediaan data ialah untuk menyediakan set data yang bersih dan berkualiti bagi perlombongan data. Antara proses yang terlibat ialah pra-pemprosesan data serta pembahagian set data latihan dan ujian.

Set data yang ingin dilombong haruslah dipilih dan di pra-proses terlebih dahulu. Dalam kajian ini, empat set data daripada pelbagai domain yang berbeza dipilih dari bank data pembelajaran *University of California Irvine* (UCI) Machine Learning (Murphy, 1997). Set data tersebut ialah Iris, Glass, Pima dan Wine. Ringkasan maklumat set data yang dipilih adalah seperti yang dipaparkan dalam Jadual 1 di bawah:

Jadual 1: Ringkasan maklumat set data

Nama	Set Data	Objek	Atribut Selanjar	Atribut Ordinal	Kelas
Iris	*Iris Plants Database*	150	4	0	3
Glass	*Glass Identification Database*	214	9	0	6
Pima	*Pima Indians Diabetes*	768	8	0	2
Wine	*Wine Recognition Data*	178	13	0	3

Kesemua set data yang dipilih hanya mengandungi atribut selanjar sahaja. Ini adalah penting kerana kajian ini hanya memfokuskan kepada pendiskritan atribut selanjar. Penggunaan atribut selanjar juga dapat menguji keberkesanan sesuatu teknik pendiskritan dalam mendiskritkan atribut selanjar. Nama atribut bagi setiap set data ditukarkan bagi memudahkan perwakilan.

3.1.1 Pra-pemprosesan Data

Pra-pemprosesan dilaksanakan ke atas set data uji kaji bagi membersihkan data dan memperbaiki kualiti data yang hendak dilombong. Dalam langkah ini, beberapa teknik pra-pemprosesan yang

sesuai diaplikasikan meliputi kategori pra-pemprosesan data seperti pembersihan data, transformasi data dan pengurangan data. Integrasi data tidak dilaksanakan kerana data hanya berasal daripada satu sumber sahaja dan tidak perlu diintegrasikan.

Dalam set data daripada UCI, kesemua set data yang dipilih tidak mempunyai nilai atribut yang hilang. Ini disebabkan oleh hampir kesemua data UCI telah menjalani pra-pemprosesan yang baik oleh penyumbangnya. Selain daripada data hilang, taburan data bagi atribut juga dikenal pasti. Oleh kerana kesemua set data yang dipilih terdiri daripada atribut selanjar, teknik pendiskritan perlu dilaksanakan kerana pengelas set kasar bertindak dengan baik dengan menggunakan atribut diskrit. Teknik pendiskritan yang dipilih bagi perbandingan ialah EF, BR dan Ent-MDLP.

Bilangan selang (*BS*) yang dihasilkan berbeza mengikut teknik pendiskritan yang digunakan. Secara keseluruhannya, hasil pendiskritan yang dihasilkan oleh teknik BR mempunyai *BS* yang paling kecil manakala teknik Ent-MDLP pula menghasilkan *BS* yang paling banyak. Teknik EF pula mempunyai bilangan selang yang statik yang ditentukan secara lalai.

3.1.2 Pembahagian Set Data Latihan Dan Ujian

Untuk memastikan model yang dibangunkan stabil dan memberikan ramalan yang konsisten, pembahagian data untuk set data latihan dan ujian dilakukan. Teknik yang digunakan dalam kajian ini ialah pengesahan silang *k*-lipatan kerana teknik ini sesuai untuk semua saiz sampel dan dapat mengekalkan integriti data. Untuk kajian ini, 10-lipatan digunakan untuk pembangunan model pengelas set kasar. Setiap set data dibahagikan kepada 10 lipatan. Setiap lipatan akan mengandungi 9 model yang dihasilkan dalam bentuk set data latihan nisbah set data ujian (set data latihan: set data ujian) seperti 90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80 dan 10:90 seperti yang ditunjukkan di dalam Jadual 2. Model yang mempunyai ketepatan yang tertinggi daripada setiap lipatan diambil untuk tujuan perbandingan.

Jadual 2: Pembahagian set data latihan dan ujian

PD_{LU}	Set Data							
	Iris (150 data)		Glass (214 data)		Pima (768 data)		Wine (178 data)	
	Lat	**Uji**	**Lat**	**Uji**	**Lat**	**Uji**	**Lat**	**Uji**
90:10	135	15	193	21	691	77	160	18
80:20	120	30	171	43	614	154	142	36
70:30	105	45	150	64	537	231	125	53
60:40	90	60	128	86	460	308	107	71
50:50	75	75	107	107	384	384	89	89
40:60	60	90	86	128	308	460	71	107
30:70	45	105	64	150	231	537	53	125
20:80	30	120	43	171	154	614	36	142
10:90	15	135	21	193	77	691	18	160

3.2 Perlombongan Data

Untuk membangunkan model bagi kajian ini, pengelas set kasar digunakan. Pengelas set kasar merupakan aplikasi kepada teori set kasar yang diperkenalkan oleh Zdzislaw Pawlak pada tahun 1982. Kesemua set data yang dipilih akan dilombong menggunakan pengelas set kasar.

3.3 *Pengujian dan Perbandingan*

Bagi setiap teknik pendiskritan, setiap data dibahagikan menggunakan pengesahan silang 10-lipatan. Setiap lipatan ini terdiri daripada 9 model yang dijana dalam bentuk set data latihan nisbah set data ujian. Set data latihan digunakan untuk membangunkan model manakala set data ujian pula digunakan bagi menentukan ketepatan model yang diperoleh daripada petua yang

dihasilkan oleh set data latihan. Teknik pendiskritan EF, BR dan Ent-MDLP dilarikan ke atas kedua-dua set data tersebut.

Setelah pengujian dilaksanakan, analisis perbandingan dijalankan. Perbandingan antara teknik EF, BR dan Ent-MDLP dibincangkan daripada segi ketepatan pengelas, bilangan petua dan panjang petua yang dihasilkan. Ketepatan pengelas adalah keupayaan model untuk mengelas atribut kelas bagi data baru dengan tepat di dalam bentuk peratusan (Mohamad Farhan, 2006). Peratus ketepatan model yang tertinggi sama ada secara individu dan purata digunakan untuk mencari model dan teknik terbaik.

Walau bagaimanapun, ketepatan pengelas yang tertinggi sahaja tidak akan menentukan sesebuah model itu adalah yang terbaik. Bilangan dan panjang petua yang dihasilkan juga mempengaruhi keputusan. Bilangan petua merujuk kepada jumlah petua yang dihasilkan daripada set data latihan. Bilangan petua yang sedikit tetapi menghasilkan ketepatan tertinggi menunjukkan tahap keberkesanan sesebuah model yang dibangunkan.

Panjang petua pula merupakan saiz petua yang dihasilkan. Panjang petua minimum dan panjang petua maksimum diambil kira dalam menentukan model yang terbaik. Model yang mempunyai panjang petua minimum dan bilangan petua yang pendek adalah model yang baik kerana bilangan petua yang banyak akan menghasilkan petua yang tidak berguna untuk pengelasan. Penghasilan petua yang pendek lebih mudah difahami.

Sekiranya dua model mempunyai ketepatan pengelas yang sama, model yang mempunyai petua yang lebih sedikit dan lebih pendek merupakan model yang lebih baik daripada model yang mempunyai petua yang banyak dan panjang. Ketiga-tiga kriteria ini saling mempengaruhi dalam menentukan model yang terbaik.

4. Hasil Pengujian dan Perbincangan

Perbandingan dilakukan daripada segi ketepatan pengelas (*K*), bilangan petua (*BP*), panjang petua (*PP*) dan bilangan selang (*BS*) yang dihasilkan. Hasil uji kaji yang dijalankan ke atas empat set data daripada UCI yang ditunjukkan dalam Jadual 3 di bawah:

Jadual 3: Hasil uji kaji

m	*EF*				*BR*				*Ent-MDLP*			
	K	*BP*	*Pmin*	*Pmak*	*K*	*BP*	*Pmin*	*Pmak*	*K*	*BP*	*Pmin*	*Pmak*
Iris	95.56	9	1	2	97.04	7	1	2	98.33	48	1	2
Glass	90.48	1249	2	6	95.24	696	1	7	85.71	432	1	5
Pima	75.52	1532	2	6	80.52	3143	2	6	76.62	61	1	7
Wine	98.11	2990	2	4	100	659	1	3	100	1021	1	6

Daripada segi K, *BR* mempunyai *K* yang lebih tinggi daripada *EF* dan *Ent-MDLP*. Jadual 4 menunjukkan perbandingan prestasi ketiga-tiga teknik pendiskritan ini. Walaupun purata keseluruhan teknik tidak jauh berbeza namun teknik *BR* telah mengatasi *EF* dan *Ent-MDLP* daripada segi *K* dalam kebanyakan model.

Jadual 4: Purata *K* ke atas *EF*, *BR* dan *Ent-MDLP*

Set data	*EF (%)*	*BR (%)*	*Ent-MDLP (%)*
Iris	97.27	99.03	96.98
Glass	73.39	84.35	76.5
Pima	74.47	75.36	71.35
Wine	97.72	99.79	92.30
Purata	**85.71**	**89.63**	**84.28**

Daripada segi *BP* yang dihasilkan pula, teknik *Ent-MDLP* didapati menghasilkan *BP* yang paling sedikit bagi setiap set data kecuali set data Iris seperti yang ditunjukkan dalam Jadual 5. Teknik

EF dan *BR* adalah setara. Perkara yang mempengaruhi jumlah *BP* ialah *BS* yang dihasilkan semasa proses pendiskritan data.

Jadual 5: Purata *BP* ke atas *EF*, *BR* dan *Ent-MDLP*

Set data	***EF***	***BR***	***Ent-MDLP***
Iris	18	15	41
Glass	1083	680	405
Pima	1762	2847	54
Wine	3011	1692	893

Daripda seg *PP*, *Pmin* dan *Pmak* petua yang dihasilkan oleh ketiga-tiga teknik adalah tidak jauh berbeza. *PP* yang paling maksimum dihasilkan oleh teknik *Ent-MDLP* dalam set data Wine iaitu lapan petua. Petua yang terlalu panjang akan menyebabkan petua tersebut terlalu khusus dan tidak sesuai digunakan ke atas data baru.

Perbincangan yang terakhir ialah daripada segi *BS* yang terhasil. Secara keseluruhannya, hasil pendiskritan yang dihasilkan oleh *BR* mempunyai *BS* yang paling kecil manakala teknik *Ent-MDLP* pula menghasilkan *BS* yang paling banyak. Teknik *EF* pula mempunyai *BS* yang statik yang ditentukan secara lalai.

Untuk memilih teknik pendiskritan terbaik berdasarkan keputusan di atas, beberapa kriteria digunakan seperti model yang mempunyai *K* tertinggi, *BP* yang sedikit, *PP* yang lebih pendek dan *BS* yang sedikit. Didapati teknik *BR* menghasilkan *K*, *PP* dan *BS* yang lebih baik berbanding teknik lain. Walaupun teknik *Ent-MDLP* menghasilkan *BP* yang baik namun teknik ini menghasilkan *BS* yang kurang baik. Kesimpulan bagi keseluruhan hasil kajian ialah teknik *BR* lebih baik daripada teknik *EF* dan *Ent-MDLP*.

5. Kesimpulan

Kajian ini dijalankan bagi memenuhi beberapa objektif. Objektif bagi kajian ini ialah mengkaji teknik pendiskritan dalam perlombongan data, membuat perbandingan prestasi di antara tiga teknik pendiskritan data yang terpilih dalam model pengelasan berdasarkan ketepatan pengelas, bilangan petua dan panjang petua dan mengaplikasikan teknik-teknik pendiskritan data yang terpilih ke atas empat set data daripada UCI *Machine Learning*.

Di akhir kajian ini, semua objektif yang dinyatakan telah tercapai. Objekif pertama telah dilaksanakan dalam kajian literatur di mana tiga teknik pendiskritan telah dikaji. Bagi objektif kedua dan ketiga, teknik BR, EF dan Ent-MDLP yang dipilih berdasarkan kajian literatur telah digunakan untuk membangunkan model perlombongan data bagi empat set data daripada UCI iaitu Iris, Glass, Pima dan Wine. Berdasarkan model-model yang diperolehi, perbandingan prestasi di antara ketiga-tiga teknik ini diukur daripada segi ketepatan pengelas, panjang petua dan bilangan petua yang dihasilkan.

Kajian ni boleh diperkembangkan dengan membandingkan teknik pendiskritan EF, BR dan Ent-MDLP dengan teknik pendiskritan yang menggunakan pendekatan penggabungan seperti Chi2, StatDisc atau lain-lain. Secara keseluruhannya, teknik pendiskritan data merupakan satu teknik yang penting dalam perlombongan data. Hasil uji kaji mendapati teknik BR merupakan teknik pendiskritan terbaik berbanding dengan teknik EF dan Ent-MDLP. Teknik BR didapati menghasilkan pendiskritan terbaik berdasarkan peratusan ketepatan pengelas yang tinggi dengan hanya menggunakan bilangan petua yang sedikit dan pendek.

6. Rujukan

Catlett, J. (1991). On changing continuous attributes into ordered discrete attributes. In Dougherty, J., Kohavi, R. & Sahami, M. (1995). Supervised and Unsupervised Discretization of Continuous Features. *Proc Twelfth International Conference on Machine Learning*, 194-202.

Cerquides, J. & López de Màntaras, R. (1997). Proposal and empirical comparison of a parallelizable distance-based discretization method. *3d Int. Conference on Knowledge Discovery and Data Mining (KDD'97),* 139-142.

Ching, J. Y., Wong, A. K. C. & Chan, K. C. C. (1995). Class-dependent discretization for inductive learning from continuous and mixed mode data. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 17(7), 641-651.

Dougherty, J., Kohavi, R. & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. *Proc Twelfth International Conference on Machine Learning*, 194-202.

Fayyad, U. M. & Irani, K. B. (1993). Multi-interval discretization of continuous valued attributes for classification learning. *Proceedings of IJCAI,* 2, 1022-1027.

Gama, J. & Pinto, C. (2006). Discretization from data streams: applications to histograms and data mining. *Proceedings of the 2006 ACM symposium on Applied computing*, 662 – 667.

Goharian, N. & Grossman, D. (2003). Data mining: data preprocessing. *Illinois Institute of Technology*. Retrieved 2 May 2006, from http://www.ir.iit.edu/~nazli/cs422/ CS422-Slides/DM-Preprocessing.pdf.

Han, J. & Kamber, M. (2001). *Data mining: concepts and techniques.* San Francisco: Morgan Kaufmann Publishers.

Kerber, R. (1992). ChiMerge: Discretization of numeric attributes. *Proceedings of the 10th National Conference on Artificial Intelligence*, 123-128.

Kotsiantis, S. B., Kanellopoulos, D. & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International Journal of Computer Science,* 1(2), 219-222.

Li, J. & Cercone, N. (2005). Empirical analysis on the geriatric care data set using rough sets theory. Technical report. School of Computer Science, University of Waterloo.

Liu, X. & Wang, H. (2005). A discretization algorithm based on a heterogeneity criterion. *IEEE Transactions on Knowledge and Data Engineering,* 17(9), 1166-1173.

Liu, H., Hussain, F., Tan, C.L. & Dash, M. (2002). Discretization: an enabling technique. *Data Mining and Knowledge Discovery,* 6, 393-423.

Mohamad Farhan Mohamad Mohsin. (2006). *Pengubahsuaian Ke Atas Algoritma Apriori Dan Perbandingannya Dengan Pengelas Kasar* (Unpublished masters's thesis). Universiti Kebangsaan Malaysia.

Murphy, P.M. (1997). *UCI repositories of machine learning and domain theories*. Retrieved 1 September 2006, from http://www.ics.uci.edu/~mlearn/MLRepository.html

Nguyen, H.S. & Skowron, A. (1997). Boolean reasoning for feature extraction problems. *International Symposium on Methodologies for Intelligent Systems 2007*, 117-126.

Nor Liyana Mohd Shuib, Azuraliza Abu Bakar, Zulaiha Ali Othman. (2009). Building a New Taxonomy for Data Discretization Techniques. *Second Conference of Data Mining and Optimization*. Universiti Kebangsaan Malaysia.

Panda, S. S. (2005). Data Visualization and Preprocessing. *GIS Training and Research Center*. Retrieved 22 September 2006, from http://giscenter.isu.edu/training/ppt/Research%20Method%20(SA)/5-6-Data%20Mining%20&%20Area%20Data%20Analysis.ppt

Pawlak, Z. (1982). *Rough sets. International Journal of Computer and Information Science***,** 11, 341-356.

Pawlak, Z. & Skowron, A. (2007). Rough sets and Boolean reasoning. *Information Science,* 177(1), 41-73.

Tay, F. & Shen, L. (2002). A modified chi2 algorithm for discretization. *IEEE Transactions of Knowledge and Data Engineering,* 14(3), 666-670.

Ventura, D. & Martinez, T. (1995). An empirical comparison of discretization methods. *Proceedings of the Tenth International Symposium on Computer and Information Sciences,* 443-450.