VOT 75056

STUDY OF CLUSTERING TRECHNIQUES IN DATA MINING FOR CLIMATE DATA

RESEARCHER:

MAHADI BIN BAHARI (HEAD)

ROZILAWATI BINTI DOLLAH @ MD. ZAIN

PM DR. MOHD NOOR BIN MD. SAP

ARYATI BINTI BAKRI

UNIVERSITI TEKNOLOGI MALAYSIA

2006

JADUAL KANDUNGAN

BAB I

PENGENALAN

1.1 Pengenalan

Penganalisaan data merupakan satu tugas yang penting bagi pihak Jabatan Perkhidmatan Kajicuaca Malaysia (JPKM), di mana hasil daripada penganalisaan ini akan digunakan untuk membuat keputusan atau peramalan cuaca pada masa akan datang. Walaubagaimanapun, pengekstrakan maklumat yang berguna daripada pangkalan data di JPKM bagi tujuan penganalisaan ini menjadi sukar ekoran pertambahan jumlah data kajicuaca yang disimpan. Pertambahan jumlah data dan jenis data yang dikumpul dan disimpan ini menyebabkan alat-alatan penganalisaan data konvensional adalah tidak mencukupi untuk mengekstrak maklumat berguna yang dikehendaki daripada pangkalan data-pangkalan data tersebut (Choenni dan Siebes, 1996).

Oleh yang demikian, perlombongan data merupakan salah satu teknik yang dapat membantu dan memudahkan pihak JPKM untuk mengekstrak data-data bermakna dan berguna sahaja daripada koleksi data yang banyak dengan cekap, cepat dan berkesan. Ini kerana perlombongan data boleh mengenalpasti tren dan corak data dalam lautan maklumat dengan melakukan penjelajahan dan analisis data secara sistematik ke atas jumlah data yang banyak dengan tujuan untuk mengekstrak maklumat-maklumat yang dikehendaki. Tambahan pula, salah satu daripada matlamat perlombongan data ialah peramalan di mana corak data bagi tujuan peramalan akan diperolehi daripada pangkalan data yang besar.

Terdapat pelbagai cara yang boleh digunakan untuk mengekstrak data yang dikehendaki daripada pangkalan data. Di antaranya ialah pengelompokan,

pengkelasan dan sebagainya, di mana hasil daripada pengekstrakan ini boleh digunakan di dalam proses membuat jangkaan atau peramalan tentang perubahan cuaca. Oleh yang demikian, kajian ini dijalankan untuk mengelompokkan data kajicuaca untuk digunakan di dalam peramalan taburan hujan. Sehubungan dengan itu, bagi melaksanakan kajian ini dua kategori algoritma pengelompokan telah dipilih untuk melakukan pengelompokan data kajicuaca iaitu pengelompokan *hierarchical* dan *partitional*.

1.2 Latarbelakang Masalah

Pihak Jabatan Perkhidmatan Kajicuaca Malaysia (JPKM) memainkan peranan yang penting di dalam memantau situasi perubahan cuaca dan mengeluarkan kenyataan, nasihat dan amaran cuaca bila keadaan memerlukan (JPKM, 2000). Oleh yang demikian, maklumat yang tepat diperlukan bagi membantu pihak JPKM membuat jangkaan atau peramalan yang terperinci tentang perubahan cuaca sebelum mengeluarkan sebarang kenyataan, nasihat atau amaran cuaca.

Sepertimana yang diketahui, pihak JPKM akan membuat pemantauan bagi semua stesen pencerapan kajicuaca di seluruh negara. Setiap stesen pencerapan ini akan memancarkan keadaan cuaca sebenar yang dicerap ke semua pejabat ramalan dan dipamerkan setiap jam (JPKM, 2000). Bilangan stesen yang banyak ini telah menyebabkan pertambahan jumlah data kajicuaca yang disimpan di JPKM. Ini kerana jumlah data kajicuaca tersebut semakin meningkat dari masa ke semasa. Antara parameter-parameter data kajicuaca ialah data taburan hujan, suhu, angin, kelembapan, sejatan dan sinaran suria (Sani, 1984). Peningkatan jumlah data ini telah menimbulkan kesukaran kepada pihak JPKM untuk melakukan proses penganalisaan data kajicuaca termasuklah bagi tujuan melakukan peramalan cuaca. Tambahan pula, tugas mencerap, memahami dan meramalkan perubahan cuaca merupakan satu tugas yang sukar dan mencabar.

Peramalan cuaca merupakan salah satu daripada masalah yang paling mencabar di dalam dunia sejak lebih setengah abad yang lalu. Ini bukan hanya disebabkan oleh nilai praktikalnya di dalam kajicuaca, tetapi ia juga merupakan

masalah peramalan siri masa yang “unbiased” di dalam penyelidikan saintifik (Liu dan Lee, 1999). Justeru itu, pelbagai kaedah peramalan telah dibangunkan sejak beberapa tahun lalu samada dengan menggunakan algoritma *hierarchical* dan *partitional*.

'

Di antara parameter-parameter kajicuaca, taburan hujan merupakan parameter yang paling sukar diramal (Liu dan Lee, 1999). Ini kerana peramalan hujan adalah merupakan satu masalah yang kompleks dan sukar kerana ia melibatkan pelbagai pembolehubah di mana ia saling berhubungkait dengan cara yang rumit. Kebanyakan perhubungannya adalah mengambarkan hubungan ruang dan dinamik yang tidak linear (Chen dan Takagi, 1993).

Pada masa ini, proses peramalan taburan hujan di Malaysia secara keseluruhannya dilakukan secara manual oleh beberapa orang peramal. Hingga ke hari ini, belum ada satu model atau formula pun yang boleh digunakan untuk melakukan peramalan hujan di Malaysia. Selepas data-data kajicuaca diperolehi sama ada dari radar, satelit ataupun lain-lain sumber seperti kapten kapal, pelantar-pelantar minyak dan sebagainya, ia akan dianalisa sekaligus. Kemudian carta dan graf akan diplot berdasarkan data-data tersebut. Seterusnya, peramalan hujan dilakukan berdasarkan kepada corak cuaca dari carta atau graf hujan dan angin. Selain daripada itu, bantuan pengalaman dan pengetahuan dari peramal-peramal juga turut menyumbang kepada peramalan hujan (JPKM, 2000). Tugas ini memakan masa yang lama dan merupakan satu tugas yang agak sukar dilakukan.

Sehubungan dengan itu, data taburan hujan telah dipilih untuk diramal di dalam kajian ini. Selain daripada itu, taburan hujan juga merupakan salah satu daripada elemen atau ciri yang penting dan menarik di dalam iklim yang sering dibincang dan diperkatakan. Ini kerana taburan hujan bagi sesuatu tempat akan memberikan pengaruh kepada bidang pertanian dan juga menimbulkan masalah-masalah kepada manusia, makhluk-makhluk lain serta tanaman-tanaman, iaitu bencana alam seperti banjir, kemarau dan sebagainya (Arakawa, 1969).

Berdasarkan kepada kajian tentang peramalan atau penganggaran taburan hujan yang telah dijalankan oleh beberapa penyelidik sebelum ini, didapati

kebanyakan kajian, contohnya kajian yang dijalankan oleh Chen dan Takagi (1993), McCullagh et. al (1999) serta Liu dan Lee (1999) telah menggunakan teknik perlombongan data, iaitu teknik rangkaian neural pintar (*Artificial Neural Network*) untuk melakukan operasi pengkelasan terhadap data taburan hujan tersebut bagi tujuan peramalan atau penganggaran taburan hujan. Walau bagaimanapun, masih terdapat kekurangan di dalam kajian yang telah dijalankan. Di antaranya ialah berlakunya tindanan di dalam set data semasa proses latihan. Sehubungan dengan itu, kajian yang mereka jalankan ini adalah tidak sesuai untuk digunakan bagi data input yang terlalu banyak.

Oleh yang demikian, kajian ini dilaksanakan untuk membantu pihak JPKM melakukan penganalisaan data kajicuaca bagi tujuan meramal taburan hujan. Sehubungan dengan itu, kajian ini memberikan penumpuan kepada operasi pengelompokan data kajicuaca di mana hasil dari pengelompokan ini akan digunakan untuk melakukan peramalan taburan hujan. Operasi pengelompokan data kajicuaca bagi tujuan peramalan taburan hujan telah dipilih di dalam kajian ini kerana terdapat banyak kajian yang melibatkan pengelompokan telah berjaya meningkatkan hasil peramalan yang dilakukan. Contohnya, kajian yang dijalankan oleh Sarjon dan Mohd Noor (2000b), mendapati pengelompokan dapat meningkatkan peramalan yang dilakukan. Selain daripada itu, Yair et.al (1999) di dalam kajiannya tentang peramalan perkataan telah menggunakan pengelompokan semantik berdasarkan metrik persamaan semantik bagi tujuan meningkatkan peramalan.

Bagi melaksanakan kajian ini, dua kategori algoritma teknik pengelompokan telah dipilih untuk melakukan pengelompokan data kajicuaca iaitu algoritma hirarki dan partitional. Sehubungan dengan itu, fokus utama kajian ini ialah untuk mengkaji algoritma yang terbaik di antara pengelompokan hirarki dan partitional di dalam mengelompokkan data kajicuaca. Kemudian, perbandingan di antara kedua-dua kategori algoritma pengelompokan ini dilakukan dengan cara meramal taburan hujan. Peramalan taburan hujan ini dilakukan dengan menggunakan rangkaian neural. Ianya bertujuan untuk menguji kelompok-kelompok yang dihasilkan oleh kedua-dua teknik ini dan seterusnya melihat teknik pengelompokan manakah yang lebih baik dan berkesan.

1.3 Penyataan Masalah

Seperti yang dinyatakan sebelum ini, proses peramalan taburan hujan di Malaysia dilakukan secara manual. Di mana berdasarkan carta hujan, angin dan maklumat cuaca yang lain, peramalan dilakukan oleh beberapa orang peramal dengan berbantukan pengalaman, kemahiran dan pengetahuan tambahan mereka. Ini merupakan satu tugas yang sukar dilakukan. Untuk membantu pihak JPKM menangani masalah yang dihadapi semasa melakukan penganalisaan data kajicuaca secara manual, maka kajian bercadang untuk melakukan penganalisaan data kajicuaca dengan cara mengelompokkan data-data tersebut kepada beberapa kelompok. Oleh yang demikian kajian ini akan menggunakan teknik perlombongan data, iaitu *hierarchical* dan *partitional* untuk mengelompokkan data kajicuaca tersebut.

Maka, penyataan masalah bagi kajian ini adalah untuk melihat :

" ***Algoritma pengelompokan manakah yang terbaik di antara hierarchical dan partitional untuk mengelompokkan data kajicuaca bagi tujuan peramalan taburan hujan.*** "

Beberapa isu telah dikenalpasti bagi menjawab persoalan kajian ini, di antaranya :

1. Apakah kategori algoritma hirarki dan partitional di dalam pengelompokan data ?

2. Bagaimana untuk melakukan peramalan hujan ?
   - Apakah algoritma yang digunakan sebelum ini?
   - Apakah kekurangan di dalam algoritma-algoritma tersebut ?

3. Kenapa pilih algoritma hirarki dan partitional untuk melakukan pengelompokan data kajicuaca?
   - Di antara algoritma-algoritma tersebut, yang manakah yang terbaik ?
   - Apakah kelebihannya ?
   - Sejauhmanakah keberkesanan algoritma tersebut di dalam melakukan pengelompokan bagi tujuan peramalan ?

1.4 Hasil Kajian

Di dalam kajian ini, peramalan taburan hujan dilakukan untuk membandingkan keberkesanan dan ketepatan di antara dua kategori algoritma pengelompokan, iaitu *hierarchical* dan *partitional*. Bagi tujuan perbandingan ini, beberapa eksperimen telah dijalankan di mana, kelompok-kelompok data kajicuaca yang dihasilkan oleh kedua-dua teknik pengelompokan ini digunakan sebagai input kepada proses peramalan taburan hujan. Daripada hasil keputusan eksperimen ini, pengukuran prestasi peramalan taburan hujan akan dibandingkan berdasarkan kepada nilai ralat RMS dan pekali korelasi yang diperolehi di dalam setiap eksperimen tersebut. Ianya bertujuan untuk menentukan tahap keberkesanan di antara algoritma tersebut. Kemudian graf diplot berdasarkan kepada nilai ralat RMS dan pekali korelasi tersebut untuk menunjukkan perbezaan dan keberkesanan di antara teknik-teknik tersebut dengan jelas. Di akhir kajian ini, keputusan penganalisaan dan perbandingan di antara kedua-dua algoritma pengelompokan akan dihasilkan.

1.5 Objektif Kajian

Objektif bagi kajian ini adalah seperti berikut :

i. Untuk mengenalpasti algoritma pengelompokan di dalam sample data kajicuaca bagi tujuan peramalan taburan hujan.
ii. Untuk menganalisa dan membandingkan algoritma manakah yang terbaik untuk melakukan pengelompokan data kajicuaca bagi tujuan peramalan taburan hujan.
iii. Untuk mendapatkan algoritma yang terbaik di antara hiraki dan partional di dalam mengelompokkan data kajicuaca bagi tujuan peramalan.

Terdapat banyak parameter data kajicuaca yang memainkan peranan di dalam melakukan peramalan taburan hujan. Di antaranya ialah suhu, kelajuan angin, tekanan, kelembapan dan sebagainya. Oleh yang demikian, kajian perlu dilakukan untuk mengenalpasti parameter manakah yang memberikan pengaruh besar ke atas peramalan taburan hujan. Salah satu cara yang boleh digunakan untuk

mengenalpasti parameter tersebut ialah dengan cara mengelompokkan parameter-parameter data kajicuaca tersebut.

1.6 Skop Kajian

Skop kajian adalah meliputi perkara-perkara berikut :

i. Kajian ini hanya melibatkan data kajicuaca (setiap jam) dari bulan September, 1993 hingga bulan Februari, 2001 yang diperolehi daripada stesen pencerapan Kluang (MPOB).
ii. Kajian ini hanya memfokus kepada isu pengelompokan data kajicuaca bagi tujuan meramal taburan hujan.
iii. Mengkaji algoritma yang manakah terbaik di antara algoritma *hierarchical* dan *partitional* bagi pengelompokan data kajicuaca.
iv. Peramalan taburan hujan dilakukan bertujuan untuk menguji algoritma pengelompokan.
v. Pengujian hanya terbatas kepada data yang telah sedia ada di dalam pangkalan data.

Di dalam proses peramalan taburan hujan, adalah tidak munasabah untuk menggunakan kesemua parameter atau atribut data kajicuaca untuk melakukan peramalan. Bagi tujuan mengenalpasti parameter manakah yang memberikan pengaruh besar kepada ketepatan hasil peramalan hujan, pengelompokan data kajicuaca perlu dilakukan. Ianya adalah untuk mendapatkan parameter atau atribut manakah yang mempunyai persamaan yang tinggi dan paling berpengaruh di dalam menghasilkan peramalan taburan hujan yang tepat.

Oleh yang demikian, skop utama kajian ini adalah memfokus kepada isu pengelompokan data kajicuaca, di mana kajian dilakukan untuk mendapatkan algoritma pengelompokan manakah yang lebih baik dan berkesan di dalam menghasilkan kelompok-kelompok data kajicuaca yang dapat membantu meningkatkan keupayaan dan prestasi peramalan taburan hujan.

1.7 Struktur Laporan

Perlaksanaan penulisan tesis ini adalah meliputi tujuh bab. Bab pertama menerangkan tentang latarbelakang masalah, penyataan masalah, objektif, skop dan hasil kajian. Bab kedua pula menerangkan tentang kajian literatur seperti data taburan hujan, peramalan cuaca, perlombongan data, operasi dan algoritma di dalam perlombongan data, algoritma *hierarchical* dan *partitional*. Manakala bab ketiga pula merangkumi rekabentuk kajian, rangkakerja operasi kajian, sumber data dan peralatan, prosidur kajian dan akhir sekali kesimpulan.

Di dalam bab keempat, proses perlaksanaan pengelompokan data kajicuaca menggunakan algoritma *partitional* diperincikan. Selain daripada itu, eksperimen peramalan taburan hujan yang dijalankan bagi tujuan pengujian algoritma pengelompokan berserta dengan keputusan peramalan turut dinyatakan. Selanjutnya, bab ini juga menerangkan proses pengelompokan data kajicuaca yang dilakukan dengan menggunakan pendekatan perlombongan data berasaskan *hierarchical*. Eksperimen yang dijalankan beserta dengan keputusan peramalan taburan hujan yang dihasilkan turut diterangkan.

Bab kelima pula memperincikan tentang perbandingan yang dilakukan berdasarkan kepada keputusan eksperimen yang dihasilkan di dalam bab keempat dan kelima. Seterusnya, bab yang terakhir iaitu bab ketujuh menerangkan tentang perbincangan, masalah yang dihadapi di dalam melakukan kajian ini, kekangan dan andaian, kelemahan dan kelebihan kajian yang dijalankan, sumbangan kajian, cadangan kajian lanjutan di masa akan datang beserta dengan kesimpulan.

BAB 2

KAJIAN LITERATUR

2.0 Pengenalan

Menurut pegawai penasihat JPKM, terdapat banyak parameter yang menjadi faktor penyebab turunnya hujan, di antaranya ialah kelembapan, arah angin, tekanan udara dan sebagainya. Faktor-faktor ini digunakan oleh peramal kajicuaca untuk membuat ramalan hujan turun (JPKM, 2000). Dari hari ke hari, jumlah data kajicuaca yang disimpan oleh pihak JPKM semakin meningkat.

Ekoran meningkatnya jumlah data yang disimpan di dalam pangkalan data-pangkalan data ini, maka adalah menjadi satu keperluan bagi mencari jalan atau cara bagaimana untuk mengekstrak pengetahuan secara automatik daripada sejumlah data yang besar. Cara untuk mengekstrak pengetahuan tersebut dipanggil melombong data. Terdapat banyak definisi bagi perlombongan data, tetapi hampir kesemuanya melibatkan pencarian atau penemuan corak-corak atau hubungan-hubungan berguna dalam pangkalan data yang besar (Yijun, 1997).

2.1 Data Taburan Hujan

Data siri masa melibatkan data-data yang dikumpulkan mengikut sela masa yang tetap (contohnya mingguan, bulanan dan lain-lain) bagi satu tempoh masa yang ditetapkan (Roselina et.al, 1998). Contoh data siri masa ialah data bagi kajicuaca seperti data taburan hujan, kelembapan, suhu, data pasaran stok, data kewangan dan sebagainya. Perlombongan data bagi data jenis siri masa ini biasanya melibatkan kajian tentang tren dan juga perkaitan di antara pembolehubah yang berlainan.

Data taburan hujan ialah bacaan jumlah hujan yang telah disukat dengan menggunakan alat khas yang dinamakan tolok hujan. Ia diukur dalam unit milimeter (mm^3) (JPKM, 2000). Data taburan hujan merupakan salah satu daripada fenomena semulajadi yang sangat berguna untuk diramal (Koskela, et. al, 1997). Ini kerana taburan hujan merupakan salah satu daripada elemen yang sukar untuk diramal (Chen dan Takagi, 1993). Menurut Liu dan Lee (1999) serta Chen dan Takagi (1993), di antara parameter-parameter kajicuaca, taburan hujan merupakan parameter yang paling sukar diramal. Ini kerana peramalan taburan hujan adalah merupakan satu masalah yang kompleks dan sukar kerana ia melibatkan pelbagai pembolehubah yang saling berhubungkait dengan cara yang rumit. Kebanyakan perhubungannya adalah menggambarkan hubungan ruang yang tidak linear.

Sehubungan dengan itu, data taburan hujan telah dipilih sebagai domain di dalam kajian ini. Selain daripada itu, ia juga dipilih kerana ia merupakan salah satu daripada elemen atau ciri yang penting dan menarik di dalam iklim yang sering dibincang dan diperkatakan. Ini kerana taburan hujan bagi sesuatu tempat akan memberikan pengaruh kepada bidang pertanian dan juga menimbulkan masalah-masalah kepada manusia, makhluk-makhluk lain serta tanaman-tanaman, di antaranya bencana alam seperti banjir, kemarau dan sebagainya (Arakawa, 1969).

Hujan juga merupakan ciri iklim tropika yang selalu berubah-ubah. Hujan mempunyai kepentingan yang besar di dalam bidang pertanian terutamanya di kawasan

tropika, iaitu kawasan-kawasan yang berada di antara Garisan Sartan dan Garisan Jadi (Nieuwolt, 1985). Oleh yang demikian, data hujan yang boleh dipercayai adalah sangat diperlukan berbanding dengan unsur-unsur atau ciri-ciri iklim yang lain di kawasan tropika.

Jumlah hujan yang diterima menentukan jenis pertanian yang dapat dijalankan dan juga jenis tanaman yang dapat ditanam di sesebuah kawasan di mana taburan hujan bermusim menetapkan masa setiap kegiatan pertanian. Perbezaan taburan hujan tahunan merupakan faktor utama yang bertanggungjawab bagi turun naiknya jumlah penghasilan tanaman (Kartasapoetra, 1986).

2.2 Peramalan Cuaca

Peramalan cuaca memainkan peranan yang penting untuk memberikan maklumat tentang cuaca dan seterusnya dapat memberikan amaran awal tentang kemungkinan berlakunya fenomena cuaca buruk, angin kencang dan laut bergelora. Maklumat-maklumat berserta amaran ini akan disebarkan melalui media massa. Ianya bertujuan untuk memberitahu orang awam dengan secepat yang mungkin tentang keadaan cuaca buruk dan juga keadaan laut yang diramalkan akan berlaku (JPKM, 2000).

Secara asasnya, tugas meramal cuaca bergantung kepada catatan keadaan udara yang merangkumi tempat-tempat yang luas di dalam masa yang singkat. Maklumat ini akan digunakan dengan pelbagai cara untuk meramal apa yang akan berlaku seterusnya. Cara ramalan tradisi dilakukan dengan menggunakan pelbagai jenis data bagi kedua-dua keadaan permukaan dan udara untuk menghasilkan peta cuaca. Dengan bantuan peta cuaca dan data, peramal akan cuba menganggarkan kelajuan dan arah angin (Yeong, 1996). Selain daripada itu, pengalaman, kemahiran dan pengetahuan tambahan juga turut memainkan peranan di dalam melakukan ramalan cuaca. Pada masa sekarang, dengan adanya kemajuan teknologi, ia membolehkan maklumat-maklumat cuaca diambil melalui satelit dan radar.

Peramalan hujan di Malaysia secara keseluruhannya dilakukan secara manual oleh beberapa orang peramal. Hingga ke hari ini, belum ada satu model atau formula pun yang boleh digunakan untuk melakukan peramalan hujan di Malaysia. Selepas data-data cuaca diperolehi sama ada dari radar, satelit ataupun lain-lain sumber seperti kapten kapal, pelantar-pelantar minyak dan sebagainya, ia akan dianalisa sekaligus. Kemudian peramalan hujan dilakukan berdasarkan kepada maklumat-maklumat tersebut dan juga corak cuaca dari carta hujan dan angin. Selain daripada itu, bantuan pengalaman dan pengetahuan dari peramal-peramal juga turut menyumbang kepada peramalan hujan (JPKM, 2000).

Secara umumnya, tempoh ramalan bagi hujan boleh dikategorikan kepada tiga iaitu;

i) ramalan jangka pendek (satu atau dua hari),
ii) ramalan jangka sederhana (lima hari), dan
iii) ramalan jangka panjang

Untuk meramalkan hujan bagi jangkamasa pendek, contohnya satu atau dua hari, peramal di Jabatan Perkhidmatan Kajicuaca Malaysia (JPKM), menggunakan data sinoptik (data seketika, iaitu sekurang-kurangnya 12 jam sebelum meramal) bagi hujan untuk melihat corak hujan, maklumat ketetapan angin seperti arah, jenis dan kelajuan angin serta maklumat tentang tekanan udara, kelembapan dan suhu. Selain daripada itu, pengetahuan dan juga pengalaman peramal juga diambil kira di dalam membuat peramalan hujan. Di Malaysia, ramalan jangka panjang tidak dilakukan dengan terperinci kerana cuaca di Malaysia sering berubah-ubah. Biasanya peramal hanya meramal secara am sahaja untuk mendapatkan gambaran secara am tentang keadaan cuaca yang bakal melanda (JPKM, 2000).

Banyak kajian yang telah dijalankan oleh penyelidik-penyelidik berkenaan dengan peramalan taburan hujan ini, di antaranya ialah Chen dan Takagi, (1993) ; McCullagh, et al (1995) ; Liu dan Lee, (1999) dan McCullagh, et al (1999). Liu dan Lee dalam kajiannya telah mencadangkan model berasaskan rangkaian untuk meramal taburan hujan. Model ini digunakan pada *multi-station* peramalan cuaca. Hasil kajiannya

mendapati bahawa model *multiple-station* adalah lebih baik daripada model *single-station* bagi tujuan peramalan taburan hujan, di mana ia telah meningkatkan peramalan taburan hujan.

Kajian yang dilakukan oleh Chen dan Takagi pula telah mencadangkan pendekatan rangkaian neural berasaskan ciri untuk melakukan peramalan taburan hujan. Beliau mendapati ketepatan pengkelasan pengagihan *intensity* taburan hujan adalah lebih tinggi berbanding dengan kaedah regrasi yang berasaskan konvensional. Manakala kajian yang dilakukan oleh McCullagh et.al pula telah melakukan pengkelasan data taburan hujan untuk dilatihkan dengan menggunakan model rangkaian neural untuk tujuan peramalan taburan hujan. Di dalam kajian ini, beliau telah membangunkan tiga rangkaian mahir bagi tujuan pengkelasan tersebut. Hasil kajiannya menunjukkan gabungan teknik-teknik mahir ini boleh digunakan untuk kebanyakan rangkaian di mana ia boleh meningkatkan ketepatan pengkelasan.

2.3 Perlombongan Data

Perlombongan data ialah suatu proses penemuan perkaitan yang berguna, corak atau bentuk dan tren melalui penapisan yang dilakukan ke atas sejumlah besar data yang disimpan di dalam gedung-gedung data menggunakan teknologi seperti teknik-teknik statistik dan juga teknik matematik (Gartner,1998).

Definisi lain bagi perlombongan data ialah proses pencarian corak dan ketetapan dalam set-set data. Pencarian ini dilakukan oleh pihak pengguna dengan menyertakan pertanyaan (*queries*) dan pencarian akan dilakukan secara automatik terhadap pangkalan data pengguna bagi kes menentukan corak. Akhirnya, apabila telah diperolehi setiap maklumat yang dikehendaki, maka ia akan dipaparkan dalam bentuk yang sesuai sama ada menggunakan graf, laporan dan lain-lain (Parsaye, 1997).

Dengan lain perkataan, perlombongan data merupakan penemuan bagi maklumat yang tidak diketahui melalui penemuan corak atau tren di dalam set data. Perlombongan data dapat mencari dan menemukan perhubungan di antara data-data dan dapat membantu meramal masa hadapan berdasarkan data–data masa lepas. Perlombongan data boleh dilaksanakan di dalam pangkalan data yang besar (Green, 1997).

Secara keseluruhannya, perlombongan data adalah suatu proses berbantukan komputer yang mengambil data-data yang berpotensi untuk digunakan daripada gudang penyimpanan data-data mentah. Enjin perisian boleh mengimbas data-data di dalam jumlah yang besar dan secara automatik akan melaporkan corak-corak data. Kaedah perlombongan data kini menjadi popular di kalangan kebanyakan para penyelidik kerana keupayaannya untuk mengekstrak maklumat yang tersembunyi di dalam sejumlah data yang besar. Ia dianggap sebagai satu alat yang pintar dan berstrategik dan amat sesuai digunakan untuk menyelesaikan pelbagai jenis masalah, contohnya peramalan, pengkelasan, pengelompokan dan sebagainya.

Kebolehannya untuk menggali maklumat daripada pangkalan data yang besar dan seterusnya menterjemahkan maklumat yang ditemui kepada pengetahuan yang berguna menjadikannya satu alat yang semakin banyak digunakan terutamanya oleh pihak pengurusan dalam sesebuah organisasi bagi membantu membuat keputusan dengan lebih bersistematik. Bagi menyelesaikan masalah-masalah tersebut, terdapat beberapa teknik perlombongan data yang boleh digunakan di antaranya ialah teknik rangkaian neural, peraturan kesatuan, pepohon keputusan dan sebagainya. Setiap teknik ini mempunyai kekuatan dan kelemahannya yang tersendiri, di mana ia digunakan untuk menyelesaikan masalah yang berlainan.

2.4 Operasi Perlombongan Data

Terdapat beberapa operasi di dalam perlombongan data yang digunakan untuk menyelesaikan masalah yang spesifik. Di antara operasi yang popular di dalam

perlombongan data ialah model peramalan, pengkelasan dan pengelompokan (Kohonen et al., 1996; Vesanto, 1997).

2.5.1 Model Peramalan

Model peramalan menggunakan model atau algoritma untuk meramal nilai pada masa hadapan. Matlamat operasi ini ialah untuk meramal medan-medan dalam pangkalan data berasaskan kepada medan yang sedia ada. Konsep model peramalan adalah mirip kepada kebolehan manusia untuk belajar melalui pemerhatian untuk memodelkan ciri-ciri terpenting bagi sesuatu fenomena. Model peramalan boleh digunakan untuk menganalisa data yang sedia ada untuk menentukan ciri-ciri tentang set data itu.

Perlombongan data mengautomasikan sesuatu proses carian maklumat di dalam pangkalan data yang besar. Contoh bagi pemasalahan dalam peramalan ialah sasaran pasaran, peramalan cuaca dan sebagainya. Perlombongan data menggunakan data yang lepas untuk mengenalpasti sasaran yang boleh memaksimumkan pulangan pelaburan. Di dalam perlombongan data, model jenis ini digunakan untuk menganalisa data di dalam pangkalan data yang sedia ada dan meramal satu nilai yang nyata di masa hadapan berdasarkan corak yang telah dikenalpasti. Sesuatu corak boleh ditakrifkan sebagai suatu peristiwa atau gabungan peristiwa-peristiwa yang berkaitan dengan data yang terdapat di dalam pangkalan data (Berson dan Smith, 1997).

Perlombongan data memainkan peranan yang penting di dalam melakukan peramalan di mana ia mempunyai keupayaan untuk menyemak ribuan pembolehubah untuk mengasingkan beberapa pembolehubah penting yang boleh diramal. Model ini dibina daripada pangkalan data yang mempunyai data-data bersejarah untuk melihat bagaimanakah corak atau peristiwa yang telah berlaku pada masa yang lepas.

Berdasarkan corak yang didapati dari data bersejarah yang dilatih, kemudian keadaan masa hadapan akan diramal.

Model peramalan dibangunkan dengan menggunakan pendekatan pembelajaran seliaan yang mempunyai dua fasa iaitu latihan dan pengujian. Latihan membina model menggunakan contoh data yang lepas dalam jumlah yang banyak di mana ia dipanggil set latihan. Manakala pengujian pula menggunakan model itu ke atas data baru yang belum lagi dilihati sebelum ini untuk menentukan ciri ketepatan model dan kebolehan fizikalnya. Bagi melaksanakan model peramalan ini, kebanyakan penyelidikan yang dijalankan adalah menggunakan teknik rangkaian neural. Teknik ini popular di dalam melakukan peramalan kerana keupayaannya untuk menyelesaikan masalah yang berkaitan dengan peramalan. Contohnya, berdasarkan kajian yang dijalankan oleh Roselina (1999), hasil kajian menunjukkan penggunaan teknik rangkaian neural dapat meningkatkan prestasi peramalan bagi data siri masa. Kajian oleh Saiful Hafizah, et. al (1997) juga mendapati teknik rangkaian neural mampu meramal tahap keberuntungan sesebuah syarikat dengan berkesan.

2.5.2 Pengkelasan

Pengkelasan merupakan pengumpulan data ke dalam kumpulan atau kelas-kelas. Ia digunakan ke atas kelas-kelas yang sejarahnya diketahui atau membangunkan kelas-kelas yang sejarahnya tidak diketahui. Pengkelasan digunakan untuk menentukan kelas spesifik untuk data bagi setiap rekod dalam pangkalan data yang sebelum ini telah dijangkakan daripada set nilai terhad yang mungkin.

Objektif utama pengkelasan ialah untuk menganalisis data input dan untuk membangunkan deskripsi dan model yang tepat untuk setiap kelas dengan menggunakan maklumat yang wujud dalam data itu sendiri. Kelas deskriptif ini akan digunakan untuk mengklasifikasikan pengujian data pada masa hadapan di mana kelas labelnya tidak diketahui. Ia juga boleh digunakan untuk membangunkan kelas data yang lebih baik. Pengkelasan mempunyai banyak aplikasinya, termasuklah dalam bidang menentukan

sasaran pasaran, pengesanan kesilapan dan dalam diagnosis perubatan. Di dalam operasi pengkelasan, data t ersimpan digunakan untuk mengetahui kedudukan data pada kumpulan yang dikenalpasti. Contoh kajian yang melibatkan pengkelasan ialah kajian yang dilakukan oleh Chen dan Takagi (1993), di mana beliau telah mengelaskan data-data taburan hujan bagi tujuan meramal taburan hujan.

2.5.3 Pengelompokan

Pengelompokan juga dikenali sebagai pengsegmenan di mana ia tidak menghasilkan medan yang spesifik untuk diramal tetapi ia menentukan sasaran data kepada beberapa subset yang sama dengan medan lain. Menurut Fayyad (1997), perangkaian ialah proses mengumpulkan objek fizikal atau abstrak yang sama ke dalam kelas-kelas untuk mengetahui corak agihan keseluruhan set-set data.

Perangkaian juga boleh ditakrifkan sebagai sekumpulan objek atau data yang dirangkaikan mengikut kriteria sama. Ia selalunya dicapai dengan mencari sekumpulan perkara data yang hampir antara satu sama lain berdasarkan kepada beberapa kriteria. Menurut TwoCrows (1999) pula, pengelompokan ialah membahagikan pangkalan data kepada beberapa kumpulan. Matlamat bagi pengelompokan ialah untuk memperolehi kumpulan yang berbeza di antara satu sama lain tetapi ahli bagi setiap kumpulan tersebut adalah hampir sama.

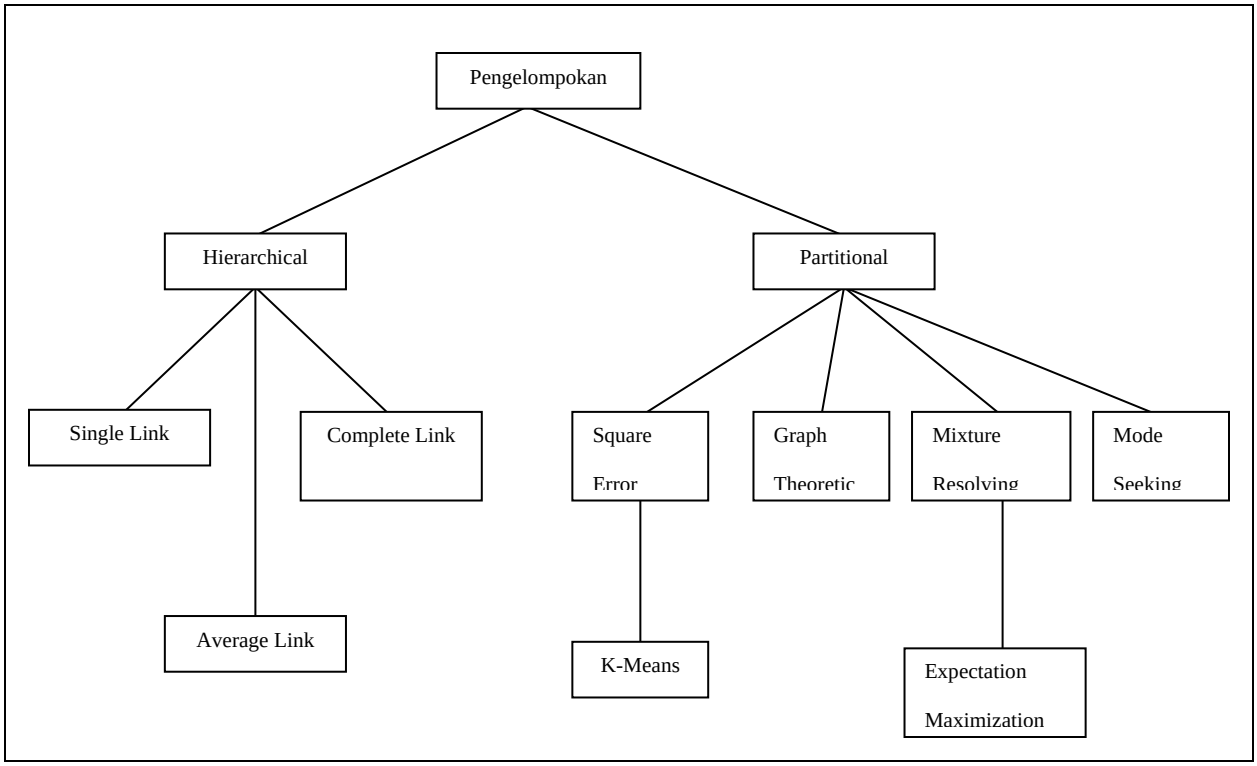
Manakala Kaski (1997) pula mentakrifkan pengelompokan merupakan kaedah untuk menghimpunkan vektor-vektor yang hampir sama atau yang mempunyai persamaan antara satu sama lain berdasarkan ciri-ciri datanya. Ia seperti merangkaikan atau menghimpunkan data-data berdasarkan persamaan ciri data tersebut. Daripada beberapa definisi yang dinyatakan di atas, maka secara keseluruhannya dapatlah dikatakan bahawa pengelompokan merupakan proses carian terhadap corak dalam pangkalan data bersejarah untuk mengelompokkan data-data yang mempunyai persamaan ke dalam satu kelompok atau kumpulan. Pengelompokan merupakan salah satu teknik utama di dalam perlombongan data di mana beberapa set entiti dibahagikan

kepada beberapa kumpulan atau subkelas yang bermakna, iaitu dipanggil kelompok. Setiap elemen yang terdapat di dalam kelompok mempunyai persamaan antara satu sama lain (atau boleh dikategorikan sebagai satu kumpulan) dan terdapat perbezaan antara satu kelompok dengan kelompok yang lain. Tujuan utama proses pengelompokan adalah untuk mengenalpasti corak sesebuah kumpulan, di mana membolehkan kita melihat persamaan serta perbezaan yang wujud antara kumpulan. Ini membolehkan andaian serta peramalan dapat dibuat berdasarkan kumpulan yang telah dikelompokan ini. Terdapat pelbagai kaedah di dalam pengelompokan di mana setiap kaedah berfungsi mengikut cara tersendiri dan mengeluarkan keputusan yang berlainan (Zait and Metsaffa, 1997).

Pengelompokan merupakan salah satu daripada masalah utama di dalam penjelajahan analisis data. Masalah pengelompokan data berlaku di dalam pengecaman corak, statistik, pembelajaran tanpa seliaan, rangkaian neural, perlombongan data, pembelajaran mesin dan pelbagai bidang saintifik yang lain (Hofmann dan Buhmann, 1997).

Operasi pengelompokan juga agak popular di kalangan penyelidik dewasa ini Banyak kajian yang telah melibatkan pengelompokan dijalankan. Contohnya kajian yang dijalankan oleh Sarjon dan Mohd Noor (2000a) mendapati, penggunaan pengelompokan dapat meningkatkan prestasi peramalan yang dilakukan. Selain daripada itu, kajian oleh Yair, et. al (1999) juga menunjukkan bahawa penggunaan pengelompokan semantik berdasarkan metrik persamaan semantik dapat meningkatkan hasil peramalan yang dilakukan. Kajian pengelompokan juga telah dijalankan oleh Sakira (1998) ; Azah (1999) dan Shahliza (1999).

Kepelbagaian teknik di dalam pengelompokan data ditunjukkan seperti pada Rajah 2.1. Rajah 2.1 menunjukkan teknik pengelompokan yang digunakan dalam proses perlombongan data. Algoritma utama di dalam teknik pengelompokan data dibahagikan kepada dua kategori iaitu *hierarchical* dan *partitional*. Seksyen seterusnya akan membincangkan mengenai kedua-dua kategori ini.



Rajah 2.1: Teknik Pengelompokan.

2.6 Pengelompokan *Partitional*

Algoritma *partitional* mempunyai kelainan berbanding penglompokan menggunakan algoritma *hierarchical* di mana ia akan membentuk data kepada *k* kelompok. Secara praktiknya, algoritma *partitional* akan dijalankan beberapa kali dengan berlainan keadaan awalan dan konfigurasi yang terbaik akan dijadikan sebagai output kepada pengelompokan yang telah dijalankan (Jain et al, 1999). Seperti yang ditunjukkan di dalam rajah 1, terdapat empat jenis algoritma pengelompokan partitional iaitu *square error*, *graph theoretic*, *mixture resolving* dan *mode seeking*. Algoritma *partitional* akan mengkategorikan data kepada beberapa kelompok yang berlainan. Ia akan mengenalpasti bilangan kelompok yang dapat dijana berdasarkan fungsi kriteria bagi tujuan mengoptimumkan data (Haldiki et al, 2001). Fungsi kriteria yang paling kerap digunakan ialah *square error* di mana setiap pengiraan jarak daripada kelompok tengah akan

dijumlahkan bagi setiap set data yang terlibat. Ia juga dikenali sebagai algoritma *square error*. Salah satu algoritma yang meminimumkan *square error* ini adalah algoritma K-Means. Seksyen seterusnya akan membincangkan mengenai algoritma Forgy dan K-Means.

The most commonly used goal in partitional clustering is to minimize a square error criterion measured by the sum of the distances from each sample in one cluster to its cluster center. Suppose that cluster k has samples $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ..., \boldsymbol{x}_{n_k})$, $\boldsymbol{x}_c$ is the center of cluster k and $e_k = \sum_{i=1}^{n_k} (\boldsymbol{x}_i - \boldsymbol{x}_c)^2$ is the variance of within the cluster. The total square error E of these K clusters is the sum of the entire cluster variations:

$$E = \sum_{j=1}^{K} e_j$$

Here the Euclidean distance is used to represent the distances between samples.

2.6.2 Algoritma Forgy

Algoritma Forgy merupakan algoritma asas yang telah dibangunkan bagi membentuk kelompok menggunakan teknik pengiraan jarak. Ia telah menjadi asas utama terutamanya dalam pembentukan algoritma K-means (Forgy, 1965).

Menurut Lin, H. (1999), algoritma Forgy adalah seperti berikut:

(i) Umpukkan k bilangan kelompok sebagai titik asas. Titik- titik tengah ini boleh dipilih secara rambang atau menggunakan kaedah heuristik bagi mendapatkan titik awalan yang baik.

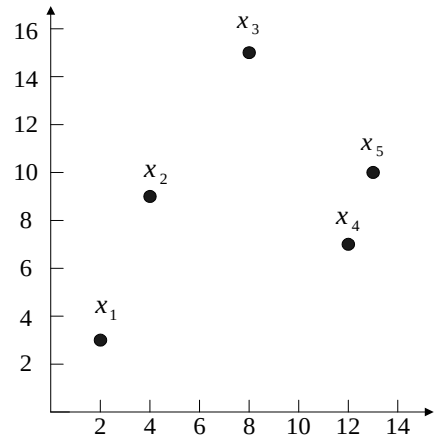
(ii) Bagi setiap sampel data, pengiraan kelompok yang terdekat akan dibuat dan pengumpukkan sampel ke dalam kelompok terdekat tersebut.

(iii) Semak sebarang perubahan kepada ahli setiap kelompok bagi setiap ulangan. Jika tiada perubahan, algoritma diberhentikan.

(iv) Kira titik tengah bagi setiap kelompok yang merupakan min vector setiap kelompok. Ulang langkah 2.

Untuk mengetahui bagaimana algoritma ini berfungsi, pertimbangkan sampel data seperti ditunjukkan pada Rajah 2.2 berikut.

$$\boldsymbol{x}_1 = (2, 3), \quad \boldsymbol{x}_2 = (4, 9), \quad \boldsymbol{x}_3 = (8, 15), \quad \boldsymbol{x}_4 = (12, 7), \quad \boldsymbol{x}_5 = (13, 10)$$



Rajah 2 : Sampel data dalam 2-dimensi

Disebabkan dua kelompok diperlukan dalam penjanaan sample ini, maka nilai $K = 2$. Perlaksanaan algoritma Forgy seperti ditunjukkan pada sampel di atas adalah seperti langkah berikut:

(i) Setkan titik awalan. Memandangkan nilai $k = 2$, secara rambang dua titik iaitu $\boldsymbol{C}_1 = \boldsymbol{x}_1$ dan $\boldsymbol{C}_2 = \boldsymbol{x}_5$, dipilih sebagai titik tengah awalan.

(ii) Bagi setiap sampel, dapatkan titik tengah bagi kelompok yang terdekat:

	Jarak ke $\boldsymbol{C}_1$	Jarak ke $\boldsymbol{C}_2$	Titik tengah Kelompok yang terdekat
$\boldsymbol{x}_1$	0	13.04	$\boldsymbol{C}_1$
$\boldsymbol{x}_2$	6.32	9.06	$\boldsymbol{C}_1$
$\boldsymbol{x}_3$	13.41	7.07	$\boldsymbol{C}_2$
$\boldsymbol{x}_4$	10.77	3.16	$\boldsymbol{C}_2$
$\boldsymbol{x}_5$	13.04	0	$\boldsymbol{C}_2$

Kemudian, bagi kelompok $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, titik tengah bagi 2 kelompok tersebut adalah $\boldsymbol{C}_1 = (\frac{2+4}{2}, \frac{3+9}{2}) = (3, 6)$ dan

$\boldsymbol{C}_2 = (\frac{8+12+13}{3}, \frac{15+7+10}{3}) = (11, 10.67)$

(iii) Dapatkan titik tengah bagi kelompok yang terdekat bagi setiap sampel tersebut sekali lagi:

	Jarak ke $\boldsymbol{C}_1$	Jarak ke $\boldsymbol{C}_2$	Titik tengah Kelompok yang terdekat
$\boldsymbol{x}_1$	3.16	11.83	$\boldsymbol{C}_1$
$\boldsymbol{x}_2$	3.16	7.2	$\boldsymbol{C}_1$
$\boldsymbol{x}_3$	10.3	5.27	$\boldsymbol{C}_2$
$\boldsymbol{x}_4$	9.06	3.8	$\boldsymbol{C}_2$
$\boldsymbol{x}_5$	10.77	2.11	$\boldsymbol{C}_2$

Selepas pengulangan semula, 2 kelompok diperolehi iaitu $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ dan $\{\boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$. Disebabkan ahli kepada kelompok yang diperolehi tidak berubah, jika dibandingkan yang pertama tdai, maka algoritma akan diberhentikan.

Seandainya bilangan sampel data terlalu besar, ianya berkemungkinan mengambil

masa yang agak lama untuk melakukan pengulangan pada pertama kali. Dalam situasi sebegini, strategi atas-talian boleh digunakan, dengan menggunakan algoritma K-Means yang akan diterangkan dalam seksyen yang berikutnya. Masalah lain yang berlaku di dalam algoritma Forgy adalah titik awalan palsu yang mungkin dipilih dan digunakan. Berdasarkan contoh di atas, jika titik awalan adalah $\boldsymbol{x}_1$ dan $\boldsymbol{x}_2$, hasil pengelompokan yang diperolehi adalah berbeza sama sekali. Kelompok akhir yang terhasil adalah $\{\boldsymbol{x}_1\}$ dan $\{\boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$ berbanding $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, dan $\{\boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$.

2.6.3 Algoritma K-Means

Algoritma K-Means adalah hampir menyerupai algoritma Forgy (MacQueen, 1967). Algoritma ini masih memerlukan bilangan kelompok yang diperlukan. Namun, ia berbeza dari segi cara untuk mendapatkan titik tengah di mana titik tengah ini akan dikira semula dengan segera apabila sample data telah bergabung di dalam kelompok terbabit.

Algoritma K-Means merupakan satu algoritma yang mudah dan kerap digunakan di dalam teknik pengelompokan kerana ia melibatkan pengiraan yang efisyen dan tidak memerlukan banyak parameter. K-Means (MacQueen, 1967) menggunakan *k* kelompok yang telah ditetapkan (k kelompok pertama sebagai centroid) dan secara berterusan akan melalui proses pengiraan titik tengah (min) sehingga sesuatu fungsi kriteria dicapai (kelompok adalah tetap). Di dalam teknik pengelompokan, pengiraan untuk membezakan di antara kelompok dilakukan menggunakan satu algoritma yang dipanggil fungsi jarak iaitu tahap persamaan atau perbezaan.

Pengukuran persamaan atau jarak merupakan tugas yang penting di dalam proses analisa kelompok di mana hampir semua teknik pengelompokan menggunakan pengiraan

matriks jarak (atau perbezaan) (Doherty et al, 2001). Algoritma K-Means juga menggunakan kaedah pengiraan ini bagi menjelaskan lagi persamaan bagi setiap corak kelompok. Matriks Jarak Euclidean merupakan salah satu matriks jarak yang kerap digunakan di dalam algoritma K-Means.

Matriks Jarak Euclidean

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

di mana $x = (x_1, x_2, ..., x_n)$ dan $y = (y_1, y_2, ..., y_n)$

Di dalam kajian ini, pengkaji telah melakukan ringkasan ke atas algoritma yang dilakukan oleh pelbagai penulis seperti yang ditunjukkan di dalam Jadual 2.1:

Jadual 2.1: Algoritma K-Means

No.	Penulis	Algoritma
1.	Kim, B. J., Kripalani, R. H., Oh, J. H., dan Moon, S. E. (2002)	(i) Wujudkan k kelas. Pilih secara rambang k corak daripada seluruh set data dan umpukkan setiap set data kepada setiap kelas. Pada fasa ini, min corak data setiap set data mengikut corak. (ii) Umpukkan setiap corak kepada set data kepada kelas di mana min yang terdekat berdasarkan pengukuran jarak $\delta_{(i,j)}$ iaitu; $\delta_{(i,j)} = \sum_{i,j=1}^{m} (X_i - X_j)^2$ (iii) Kira nilai min yang baru bagi setiap kelas. (iv) Ulang langkah (ii) dan semak jika berlaku sebarang

			perubahan corak pada kelas. Jika ya, ulang langkah (iii) dan (iv).
2.	Al-Harbi, S. H., Rayward-Smith, V. J. (2003)	(i)	Pemilihan secara rambang k kelompok, $C_i, 1 \le i \le k$ dan pengiraan centroid bagi setiap kelompok, $\hat{c}_i$.
		(ii)	Kira jarak antara objek dan centroid bagi setiap kelompok.
		(iii)	Umpukkan semula objek pada setiap kelompok.
		(iv)	Ubah centroid bagi setiap kelompok daripada yang telah dibuang dan setiap objek yang telah diumpukkan.
		(v)	Langkah (ii) dan (iv) diulang sehingga kelompok stabil.
3.	Phillips, S. J. (2002)	(i)	Anggapkan $u_1,...,u_k$ menjadi min setiap kelas. Umpukkan setiap titik $p \in P$ kepada kelas C_j yang meminimumkan $d(p,u_j)$.
		(ii)	Kira semula min; bagi setiap $j \in \{1...k\}$, set u_j yang menjadi min bagi setiap titik yang diumpukkan C_j di dalam langkah (i).
4.	Wan, S. J., Wong, S. K. M., dan Prusinkiewicz, P. (1988)	(i)	Pilih k kelompok awalan.
		(ii)	K kelompok dibentuk dengan mengumpukkan setiap data kepada kelompok yang terdekat.
		(iii)	Centroid bagi setiap k kelompok menjadi titik tengah yang baru bagi kelompok.

		(iv)	Langkah akan diulang sehingga kelompok baru yang dibentuk sama dengan sebelumnya.
5.	Pena, J. M., Lozana, J. A., dan Larranaga, P. (1999)	(i)	.Pilih pembahagian awalan setiap data kepada k kelompok $\{C_1,...,C_k\}$.
		(ii)	Kira centroids $\overline{w}_i = \frac{1}{K_i}\sum_{j=1}^{K_i} w_{ij,} \quad i = 1,...,K$. Bagi setiap w_i di dalam data dan mengikut susunan objek, Umpukkan objek w_i kepada centroid terdekat, $w_i \in C_s$ dipindahkan daripada C_s kepada C_t jika $\|w_i - \overline{w}_t\| \le \|w_i - \overline{w}_j\|$ bagi setiap $j = 1,...,K, j \ne s$.
		(iii)	Kira semula centroids bagi setiap kelompok C_s dan C_t.
		(iv)	Jika data setiap kelompok stabil maka proses diberhentikan. Jika tidak ulang langkah (iii).
6.	Cheung, Y. (2003)	(i)	Umpukkan k kelompok awalan, dan kira nilai asas $\{m_j\}_{j=1}^{k}$. Jika $j = \arg\min_{1\le r\le k}\|x_t - m_r\|^2$;
		(ii)	Diberi input x_t, kira $I(j \mid x_t)\{^{1 \quad \text{If } j=\arg\min_{1\le r\le k}\|x_t - m_r\|;}_{0 \quad otherwise}$
		(iii)	Kemaskini nilai *winning seed point* m_w ,melalui $m_w^{new} = m_w^{old} + \eta(x_t - m_w^{old})$, Di mana η merupakan *small positive learning rate.*
		(iv)	Ulang langkah (ii) dan (iii) bagi setiap input.
7.	Bandyopadhyay, S., dan Maulik,	(i)	Pilih k kelompok awalan $z_1, z_2,..., z_K$ secara rambang daripada n data $\{x_1, x_2,..., x_n\}$.

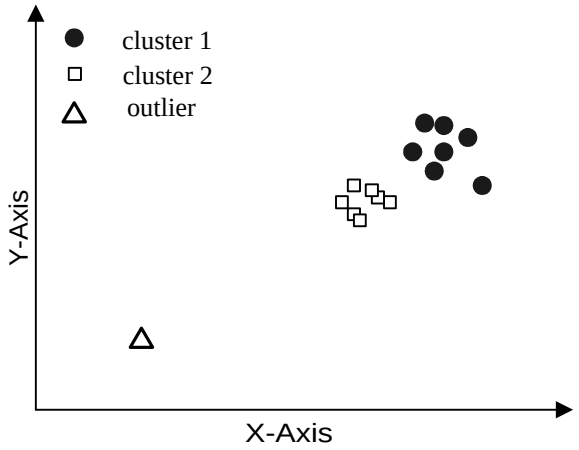
	U. (2002)	(ii) Umpukkan data $x_i, i = 1,2,...,n$ kepada kelompok C_j, $j \in \{1,2,...,K\}$ jika $\|x_i - z_j\| \leq \|x_i - z_p\|$, $p = 1,2,...,K$, dan $j \neq p$. (iii) Kira kelompok $z_1^*, z_2^*,..., z_K^*$, seperti berikut: $z_i^* = \frac{1}{n_i} \sum_{x_j \in c_i} x_j$, $i = 1,2,...,K$, Di mana n_i merupakan elemen bagi kelompok C_i. Jika $z_i^* = z_i \forall i = 1,2,...,K$ maka proses diberhentikan. (iv) Selain itu ulang langkah (ii).
8.	Smith, K. A., dan Ng, A. (2003)	(i) Nilai awalkan k kelompok sebagai kelompok tengah (guna k kelompok pertama sebagai asas). (ii) Umpukkan setiap data kepada kelompoknya yang terhampir (pengiraan daripada kelompok tengah). Ini dilakukan oleh setiap data x dan pengiraan persamaan (jarak) d melalui input ini kepada berat, w bagi setiap kelompok tengah, j. Kelompok tengah yang terhampir dengan set data x ialah kelompok tengah dengan jarak minimum dengan data x. $d_j = \|x - w_j\| = \sqrt{\sum_{i=1}^{n} (x_i - w_{ij})}$ (iii) Kira semula titik tengah bagi setiap kelompok sebagai centroid bagi setiap set data dalam setiap kelompok. Centroid $\underset{\sim}{c}$ dikira seperti berikut:

		ρ $c =< w_1^c, w_2^c, ..., w_n^c >$ Di mana $w_1^c = \frac{\sum_{j \in c} u_i^j}{N^c}$ Di mana : N^c merupakan bilangan data di dalam kelompok. (iv) Jika kelompok tengah baru adalah berlainan dengan sebelumnya, ulang langkah (ii). Jika tidak, proses diberhentikan.

Berdasarkan Jadual 2.1 di atas, dapat disimpulkan bahawa tujuan utama K-Means ialah mengenalpasti k kelompok sebagai centroid dan mengumpukkan data kepada centroid yang terhampir (sama). Pada tahap ini, pengiraan semula k kelompok baru berdasarkan hasil sebelumnya. Selepas mendapat *k* kelompok yang baru, pengiraan kepada *centroid* yang terdekat perlu dilakukan ke atas semua set data. Proses semakan akan dilakukan bagi memastikan setiap set data menepati persamaan dengan *centroid*. Proses ini akan berulang sehingga tidak berlaku perubahan ke atas lokasi centroid atau dengan kata lain, tidak ada perbezaan lagi antara set data dengan centroid. Di dalam kajian ini, penggunaan algoritma K-Means diambil daripada penulis Smith, K. A., and Ng, A. (2003).

Algoritma Forgy dan K-Means masing-masing mempunyai kelebihan seperti mudah untuk dilaksanakan dan menemui bilangan pengulangan yang minima. Namun apabila sample data mempunyai outlier yang kedudukannya agak jauh daripada data-data yang lain (contoh seperti pada Rajah 2.3), ianya akan mempengaruhi terhadap keputusan pengelompokan yang dihasilkan. Isu lain yang berlaku pada K-means, sebagaimana terdapat di dalam algoritma Forgy, adalah bagaimana untuk memlih set titik awalan, kerana pemilihan titik awalan yang kurang sesuai akan menghasilkan pengelompokan yang tidak begitu baik. Bagi mengatasi masalah ini, salah satu penyelesaiannya adalah

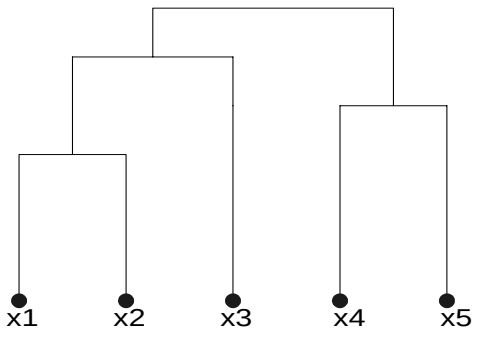
dengan mengendalikan seberapa banyak larian terhadap algoritma pengelompokan, dengan setiap satu dimulai dengan carian titik tengah yang berlainan.



Rajah 2.3: *Outlier* di dalam sempel data

2.7 Pengelompokan *Hierarchical*

Pengelompokan *hierarchical* merupakan satu jujukan *partitio*n di mana setiap *partition* disarangkan terhadap *partition* berikutnya secara berjujukan. Struktur pepohon atau dendrogram, seperti yang ditunjukkan di dalam Rajah 2.4 boleh menggambarkan pengelompokan *hierarchical* berlaku. Pengelompokan hierarchical mudah dilakukan apabila sample data telah diasingkan terlabih dahulu.



Rajah 2.4: Dendrogram bagi pengelompokan *hierarchical*

Pengelompokan Hierarchical dibahagikan kepada dua algoritma iaitu *Agglomerative* dan *Divisive*. Dalam algoritma *Agglomerative*, kelompok dihasilkan melalui pembentukan susunan skima pengelompokan dengan mengurangkan jumlah kelompok pada setiap langkah pengelompokan. Kelompok yang dihasilkan diperolehi daripada gabungan kelompok-kelompok yang terhampir (sama) kepada satu kelompok. Manakala dalam algoritma *Divisive*, kelompok dijana dengan menghasilkan susunan skima pengelompokan melalui penambahan bilangan kelompok bagi setiap langkah pengelompokan. Kelompok yang dihasilkan adalah dengan memisahkan kelompok kepada dua (Haldiki et al, 2001).

Namun begitu, skop kajian ini hanya tertumpu kepada algoritma *Agglomerative* sebagai kaedah pengelompokan yang akan digunapakai ke atas set data kajian. Algoritma *Agglomerative* dibahagikan kepada tiga jenis iaitu algoritma *Single Link*, algoritma *Average Link* dan algoritma *Complete Link*. Setiap algoritma ini mempunyai perbezaan dari segi pengiraan jarak di antara kelompok-kelompok bagi pembentukan sesuatu kelompok (Lin, 1999).

Menurut Lin (1999), algoritma *Agglomerative* dimulakan dengan n kelompok dan setiap kelompok mengandungi satu sampel. Selepas itu, setiap kelompok akan digabungkan berdasarkan persamaan di antara kelompok (jarak terdekat) sehingga kelompok dikelompokkan mengikut yang dikehendaki atau sehingga pembentukan satu kelompok. Berikut adalah langkah umum di dalam *Agglomerative*:

(i) Bermula dengan n kelompok dan satu sampel setiap kelompok.
(ii) Ulang langkah 3 sehingga kelompok yang dikehendaki atau satu kelompok.
(iii) Cari persamaan antara kelompok dan gabungkan kelompok yang sama.

Pengiraan persamaan atau jarak di antara kelompok ditentukan dengan menentukan pasangan kelompok yang sepatutnya dikelompokan. Pengiraan ini berbeza mengikut jenis algoritma *Agglomerative* yang digunakan. Cara yang piawai untuk mengira persamaan di

antara dua titik adalah dengan mengenalpasti fungsi jarak yang mengukur jarak di antara keduanya. Jarak Euclidean di antara dua titik adalah $\boldsymbol{a} = (a_1, a_2, ..., a_n)$ dan $\boldsymbol{b} = (b_1, b_2, ..., b_n)$ adalah:

$$d(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2} \ .$$

Seksyen berikut akan menerangkan algoritma-algoritma di dalam pengkelasan *Agglomerative*.

2.7.2 Algoritma *Single Link*

Bagi algoritma *Single Link*, pengiraan jarak di antara kelompok dilakukan dengan mencari jarak yang terdekat di antara sampel dalam satu kelompok dan sampel di dalam kelompok yang lain.

Formula jarak algoritma *Single Link*:

$$d(C_i, C_j) = \min_{a \in C_i, b \in C_j} d(a, b)$$

2.7.3 Algoritma *Average Link*

Manakala di dalam algoritma *Average Link*, pengiraan jarak di antara kelompok dilakukan dengan menentukan jarak purata di antara sampel di dalam kelompok atau sampel dengan kelompok lain.

Formula jarak algoritma *Average Link*:

$$d(C_i, C_j) = \frac{1}{n_{c_i} n_{c_j}} \sum_{a \in C_i, b \in C_j} d(a, b)$$

Di mana n_{c_i} merupakan jumlah kelompok bagi kelompok C_i.

2.7.4 Algoritma *Complete Link*

Pengiraan jarak di dalam algoritma *Complete Link* pula diperolehi dengan menentukan jarak yang terpanjang di antara sampel di dalam satu kelompok dan sampel di dalam kelompok yang lain.

Formula jarak algoritma *Complete Link*:

$$d(C_i, C_j) = \max_{a \in C_i, b \in C_j} d(a,b)$$

2.7.5 Contoh Perlaksanaan Algoritma Agglomerative

Berikut merupakan contoh numerik yang akan digunakan ke atas algoritma Agglomerative. Andaikan terdapat 5 sampel data ($x_1 = 2$), ($x_2 = 11$), ($x_3 = 0$), ($x_4 = 6$), dan ($x_5 = -4$). Pada awalnya, setiap sampel x_i menyatakan suatu kelompuk. Dengan menggunakan matriks D, jarak Euclidean di antara sampel adalah:

$$D = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{bmatrix} - & 9 & 2 & 4 & 6 \\ 9 & - & 11 & 5 & 15 \\ 2 & 11 & - & 6 & 4 \\ 4 & 5 & 6 & - & 10 \\ 6 & 15 & 4 & 10 & - \end{bmatrix} \end{array},$$

di mana D_{ij} menandakan jarak di antara x_i dan x_j.

Berdasarkan algoritma *single-link* agglomerative ke atas set data numerik di atas, ianya menggunakan langkah-langkah berikut:

i. Gabungkan x_1 dan x_3 ($d_{\min} = d(x_1, x_3) = 2$), dan matriks jarak D dikemaskini menjadi:

$$\begin{array}{c} \\ \{x_1, x_3\} \\ x_2 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{cccc} \{x_1, x_3\} & x_2 & x_4 & x_5 \end{array} \\ \begin{bmatrix} - & 9 & 4 & 4 \\ 9 & - & 5 & 15 \\ 4 & 5 & - & 10 \\ 4 & 15 & 10 & - \end{bmatrix} \end{array}$$

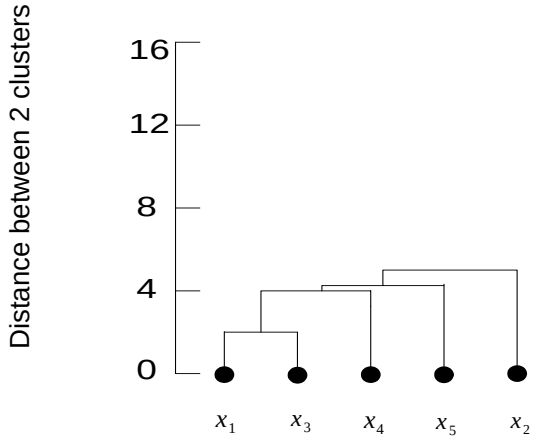
ii. Gabungkan $\{x_1, x_3\}$ dan x_4 $(d_{\min} = d(\{x_1, x_3\}, x_4) = 4)$, dan matiks jarak dikemaskini menjadi:

$$\begin{array}{c} \\ \{x_1, x_3, x_4\} \\ x_2 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccc} \{x_1, x_3, x_4\} & x_2 & x_5 \end{array} \\ \begin{bmatrix} - & 5 & 4 \\ 5 & - & 15 \\ 4 & 15 & - \end{bmatrix} \end{array}$$

iii. Gabungkan $\{x_1, x_3, x_4\}$ dan x_5 $(d_{\min} = d(\{x_1, x_3, x_4\}, x_5) = 4)$, dan matriks jarak dikemaskini menjadi:

$$\begin{array}{c} \\ \{x_1, x_3, x_4, x_5\} \\ x_2 \end{array} \begin{array}{c} \begin{array}{cc} \{x_1, x_3, x_4, x_5\} & x_2 \end{array} \\ \begin{bmatrix} - & 5 \\ 5 & - \end{bmatrix} \end{array}$$

iv. Akhirnya, semua sampel digabungkan menjadi satu kelompok $\{x_1, x_2, x_3, x_4, x_5\}$, seperti ditunjukkan pada Rajah 2.5.



Rajah 2.5: Dendrogram yang diperolehi daripada algoritma *single-link*

Langkah-langkah melalui penggunaan algoritma *complete-link* ke atas set data numerik di atas adalah:

i. Gabungkan x_1 dan x_3 ($d_{\min} = d(x_1, x_3) = 2$), dan matriks jarak D dikemaskini menjadi:

$$\begin{array}{c} \\ \{x_1, x_3\} \\ x_2 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{cccc} \{x_1, x_3\} & x_2 & x_4 & x_5 \end{array} \\ \begin{bmatrix} - & 11 & 6 & 6 \\ 11 & - & 5 & 15 \\ 6 & 5 & - & 10 \\ 6 & 15 & 10 & - \end{bmatrix} \end{array}$$

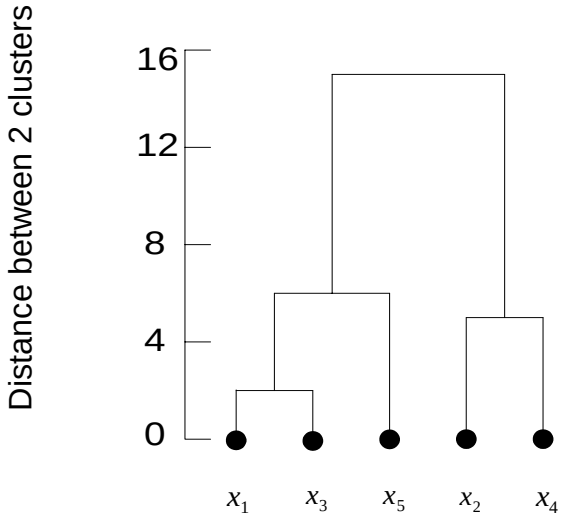
ii. Gabungkan x_2 dan x_4 ($d_{\min} = d(x_2, x_4) = 5$), dan matriks jarak dikemaskini menjadi:

$$\begin{array}{c} \\ \{x_1, x_3\} \\ \{x_2, x_4\} \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccc} \{x_1, x_3\} & \{x_2, x_4\} & x_5 \end{array} \\ \begin{bmatrix} - & 11 & 6 \\ 11 & - & 15 \\ 6 & 15 & - \end{bmatrix} \end{array}$$

iii. Gabungkan $\{x_1, x_3\}$ dan x_5 ($d_{\min} = d(\{x_1, x_3\}, x_5) = 6$), dan matriks jarak dikemaskini menjadi:

$$\begin{array}{c} \\ \{x_1, x_3, x_5\} \\ \{x_2, x_4\} \end{array} \begin{array}{c} \begin{array}{cc} \{x_1, x_3, x_5\} & \{x_2, x_4\} \end{array} \\ \begin{bmatrix} - & 15 \\ 15 & - \end{bmatrix} \end{array}$$

iv. Akhirnya, semua sample dibagungkan menjadi satu kelompok $\{x_1, x_2, x_3, x_4, x_5\}$, seperti terdapat pada Rajah 2.6.



Rajah 2.6: Dendrogram yang diperolehi daripada algoritma *complete-link*

Manakala, melalui penggunaan algoritma agglomerative average-link pula, langkah-langkahnya adalah seperti berikut:

i. Gabungkan x_1 dan x_3 ($d_{\min} = d(x_1, x_3) = 2$), dan matriks jarak D yang dikemaskini adalah:

$$\begin{array}{c} \\ \{x_1, x_3\} \\ x_2 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{cccc} \{x_1, x_3\} & x_2 & x_4 & x_5 \end{array} \\ \begin{bmatrix} - & 10 & 5 & 5 \\ 10 & - & 5 & 15 \\ 5 & 5 & - & 10 \\ 5 & 15 & 10 & - \end{bmatrix} \end{array}$$

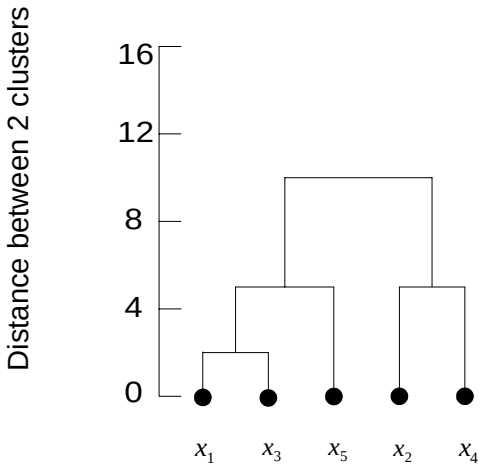
ii. Gabungkan x_2 dan x_4 ($d_{\min} = d(x_2, x_4) = 5$), dan matriks jarak dikemaskini menjadi:

$$\begin{array}{c} \\ \{x_1, x_3\} \\ \{x_2, x_4\} \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccc} \{x_1, x_3\} & \{x_2, x_4\} & x_5 \end{array} \\ \begin{bmatrix} - & 7.5 & 5 \\ 7.5 & - & 12.5 \\ 5 & 12.5 & - \end{bmatrix} \end{array}$$

iii. Gabungkan $\{x_1, x_3\}$ dan $x_5 (d_{\min} = d(\{x_1, x_3\}, x_5) = 5)$, dan matriks jarak dikemaskini menjadi seperti berikut:

$$\begin{array}{c} \\ \{x_1, x_3, x_5\} \\ \{x_2, x_4\} \end{array} \begin{array}{c} \begin{array}{cc} \{x_1, x_3, x_5\} & \{x_2, x_4\} \end{array} \\ \begin{bmatrix} - & 10 \\ 10 & - \end{bmatrix} \end{array}$$

iv. Akhirnya, semua sample digabungkan menjadi satu kelompok $\{x_1, x_2, x_3, x_4, x_5\}$, seperti pada Rajah 2.7 berikut.



Rajah 2.7: Dendrogram yang diperolehi daripada algoritma *average-link*

2.8 Ringkasan

Penulisan bab dua telah menerangkan definisi bagi perlombongan data, peramalan data siri, peramalan cuaca dan juga data taburan hujan. Di dalam perlombongan data, terdapat beberapa operasi yang boleh dilakukan untuk mengekstrak corak atau maklumat yang berguna bagi tujuan membuat keputusan dan juga menyelesaikan masalah. Operasi ini boleh dikategorikan kepada model peramalan, pengkelasan, pengelompokan dan sebagainya.

Perlombongan data juga mempunyai beberapa kategori algoritma yang digunakan sebagai alat untuk menyokong di dalam mengekstrak maklumat dan juga menemukan pengetahuan baru. Kategori algoritma yang ada dalam perlombongan data adalah *hierarchical* dan juga *partitional*. Di antara algoritma di dalam kategori *hierarchical* adalah *Agglomerative* iaitu *Single Link, Average Link* dan juga *Complete Link*. Manakala algoritma yang terdapat di dalam kategori partitional pula adalah *Forgy* dan juga *K-Means.*

BAB 3

METODOLOGI KAJIAN

3.1 Pengenalan

Bab ini membincangkan pendekatan dan metodologi kajian yang digunakan di dalam melakukan pengelompokan data kajicuaca dengan menggunakan pengelompokan partitional (algoritma Forgy dan K-Means) dan pengelompokan hierarchical (algoritma agglomerative) serta pengujian peramalan data taburan hujan yang telah dikelompokkan dengan menggunakan model rangkaian neural.

3.2 Rekabentuk Kajian

Kajian ini merupakan gabungan di antara kajian empirikal dan kajian perbandingan yang menggunakan kaedah eksperimen. Eksperimen dilaksanakan dengan menggunakan model rangkaian neural untuk melakukan peramalan data taburan hujan eksperimen yang dijalankan ini adalah bertujuan untuk menguji keberkesanan di antara teknik pengelompokan statistik dan peraturan kesatuan. Pengujian ini dilakukan menerusi perlaksanaan beberapa eksperimen. Secara amnya, kajian ini dikategorikan kepada tiga bahagian, iaitu :

a) Pengelompokan data kajicuaca menggunakan pengelompokan *partitional* (algoritma Forgy dan K-Means).

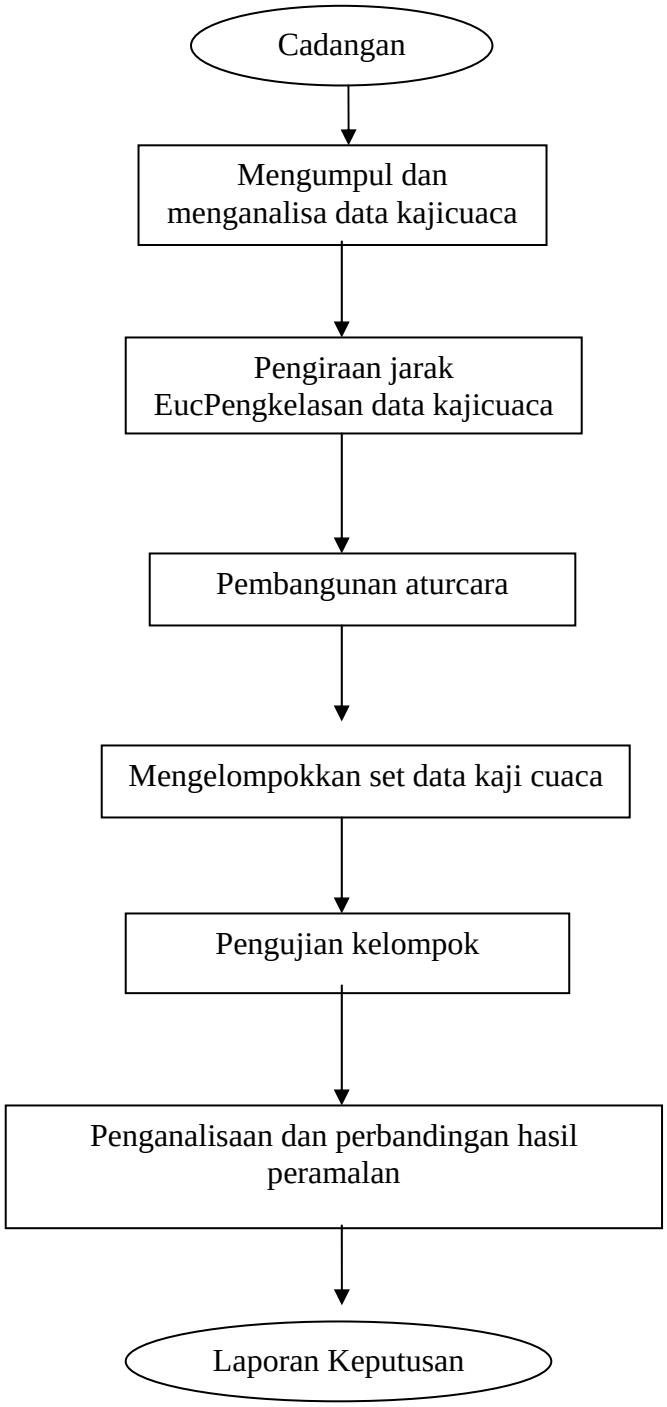
b) Pengelompokan data kajicuaca menggunakan pengelompokan *hierarchical* (algoritma agglomerative).
c) Pengujian kelompok data kajicuaca yang dihasilkan di atas dengan melakukan peramalan taburan hujan dengan menggunakan model rangkaian neural.

Kemudian penganalisaan dan perbandingan akan dibuat ke atas hasil peramalan taburan hujan yang dilakukan terhadap data kajicuaca yang telah dikelompokkan dengan menggunakan kedua-dua kaedah pengelompokan tersebut. Tujuan perbandingan ini dilakukan ialah untuk melihat dan menilai sejauhmanakah keberkesanan dan ketepatan hasil peramalan data taburan hujan yang dikelompokkan dengan menggunakan pengelompokan *partitional* dan pengelompokan *hierarchical.*

Penentuan kaedah pengelompokan yang terbaik dapat ditentukan melalui nilai ralat RMS dan juga pekali korelasi yang dihasilkan dalam pengujian peramalan taburan hujan bagi setiap eksperimen yang dijalankan. Ralat ramalan merupakan perbezaan di antara nilai sebenar dengan nilai yang diramalkan oleh model rangkaian neural. Ianya diukur dengan menggunakan formula Ralat Min Punca Kuasadua (RMS). Manakala pekali korelasi pula merupakan nilai yang mengukur kekuatan dan arah hubungan dua pembolehubah, di mana hubungan linear yang kuat menyebabkan kedua-dua pembolehubah tersebut sesuai digunakan untuk membuat peramalan.

3.3 Rangkakerja Operasi Kajian

Kajian yang dilakukan adalah mengikut rangkakerja operasi seperti yang ditunjukkan oleh Rajah 3.1. Aktiviti-aktiviti yang terlibat di dalam perlaksanaan kajian ini adalah seperti berikut :-



Rajah 3.1: Rangkakerja operasi kajian

Kajian ini dilaksanakan mengikut beberapa aktiviti. Antara aktiviti-aktiviti yang terlibat di dalam perlaksanaan kajian ini adalah seperti berikut :-

i) **Mengumpul dan menganalisa data kajicuaca** : perlaksanaan kajian ini melibatkan penggunaan 500 set data kajicuaca yang diperolehi daripada Jabatan Perkhidmatan Kajicuaca Malaysia bermula dari 1 September 2000 sehingga 21 September 2000. Ia mengandungi tujuh atribut utama iaitu *windvane, humidity, energy, temp, tension, radiation, windspeed* dan *rainfall*.

ii) **Pengiraan jarak *Euclidean*** : jarak *Euclidean* digunakan untuk mengira persamaan di antara kelompok bagi membolehkan kelompok dapat dikelompokkan dalam kelompok yang sama. Formula untuk pengiraan jarak *Euclidean* :

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

di mana $x = (x_1, x_2, ..., x_n)$ dan $y = (y_1, y_2, ..., y_n)$

iii) **Pembangunan aturcara** : selepas formula pengiraan kesamaan antara kelompok dikenalpasti, algoritma pengelompokan Forgy dan K-Means (pengelompokan *partitional*) dan algoritma pengelompokan agglomerative iaitu *single link, average link* dan *complete link* (pengelompokan *hierarchical*) dibangunkan menggunakan bahasa pengaturcaraan C.

iv) **Pengelompokan set data kajicuaca** : melalui aturcara yang telah dibangunkan, set data kajicuaca akan dikelompokkan kepada beberapa kelompok iaitu 2,3,4,5 dan 6 kelompok. Ini dilakukan kepada beberapa set data yang terdiri daripada 100 set, 200 set, 300 set, 400 set dan 500 set data kajicuaca. Hasil daripada langkah ini ialah elemen kelompok yang

terhasil, bilangan pusingan algoritma, nilai purata dan nilai sisihan piawai.

v) **Pengujian kelompok** : setelah pengelompokan dijalankan ke atas set data, setiap kelompok akan diuji menggunakan Pakej NeuNet Pro 2.3 bagi tujuan peramalan atribut taburan hujan. Ia melibatkan beberapa eksperimen berdasarkan kelompok-kelompok data kajicuaca yang sama dan kelompok data kajicuaca yang berlainan. Pengujian kelompok dilakukan untuk membuat penganalisaan dan perbandingan prestasi peramalan taburan hujan di antara set-set eksperimen yang dijalankan.

vi) **Penganalisaan dan perbandingan hasil** : melalui pengujian kelompok tadi akan menghasilkan satu keputusan peramalan di antara kelompok-kelompok data kajicuaca yang sama dan kelompok-kelompok data kajicuaca yang berlainan. Penganalisaan akan dilakukan ke atas hasil pengujian untuk menentukan ketepatan teknik pengelompokan menggunakan algoritma pengelompokan hirarki. Penganalisaan ini dilakukan dengan melihat nilai perbezaan di antara nilai sebenar taburan hujan serta nilai ramalan taburan hujan yang dihasilkan. Pengiraan peratus ketepatan juga dilakukan berdasarkan bilangan data yang diramal dan data sebenar. Selain itu juga elemen-elemen yang akan digunakan sebagai perbandingan ialah nilai ralat min punca kuasa dua (RMS) dan pekali korelasi bagi menentukan keberkesanan teknik pengelompokan hirarki.

3.4 Sumber Data dan Peralatan

Sumber data dan peralatan yang dipilih mestilah bersesuaian dengan kajian ini. Oleh yang demikian, untuk mengenalpasti apakah yang diperlukan untuk melaksanakan kajian ini, pemerhatian dan kajian dibuat ke atas kajian yang sedia ada daripada artikel-

artikel yang berkaitan. Perkhidmatan internet juga digunakan di dalam membantu untuk mendapatkan maklumat dan juga pandangan bagi kajian ini.

Memandangkan skop kajian ini ialah untuk membuat perbandingan di antara teknik peraturan kesatuan dengan kaedah statistik di dalam mengelompokkan data berstruktur, maka data kajicuaca telah dipilih untuk digunakan sebagai sumber data. Data yang akan digunakan di dalam kajian ini ialah data kajicuaca dari bulan September, 1993 hingga bulan Februari, 2001 yang diperolehi daripada stesen pencerapan Kluang (MPOB). Data-data ini telah dianggap bersih dan bebas daripada hingar. Sumber data bagi kajian ini boleh dirujuk pada lampiran B.

3.5 Prosidur Kajian

Prosidur-prosidur bagi melaksanakan kajian ini adalah seperti berikut:

i) Analisa keperluan,
ii) Perolehan dan pemprosesan data,
iii) Pengelompokan data kajicuaca,
iv) Peramalan taburan hujan,
v) Penganalisaan dan Penilaian, dan
vi) Perbandingan

3.5.1 Analisa Keperluan

Di dalam fasa analisa keperluan ini, segala maklumat berkaitan yang diperolehi akan dikaji dan diteliti bagi tujuan pemahaman tentang apakah masalah, objektif dan skop kajian. Pemahaman tentang teknik peraturan kesatuan dan kaedah statistik untuk mengelompokkan data kajicuaca juga diteliti. Selain daripada itu, pakej NeuNet Pro

juga dikaji untuk digunakan dalam meramalkan data taburan hujan bagi tujuan perbandingan. Cara penganalisaan dan peramalan taburan hujan yang dilakukan oleh pihak Jabatan Perkhidmatan Kajicuaca Malaysia (JPKM) juga turut diteliti untuk mendapat pemahaman yang lebih tentang proses peramalan taburan hujan yang dilakukan.

3.5.2 Perolehan dan Pemprosesan Data

Data kajicuaca yang diperolehi ialah data kajicuaca yang dicerap di stesen pencerapan Kluang (MPOB). Data ini adalah data kajicuaca mengikut jam, iaitu daripada bulan September, 1993 hingga bulan Februari, 2001. Kemudian, satu pangkalan data dibangunkan bagi tujuan penyimpanan data-data kajicuaca dengan menggunakan perisian Microsoft Access.

3.5.3 Pengelompokan Data Kajicuaca

Dalam prosidur ini, data kajicuaca akan dikelompokkan menggunakan dua cara iaitu;

i) menggunakan algoritma pengelompokan partitional (Forgy dan K-Means), dan

ii) menggunakan algoritma pengelompokan hierarchical(Agglomerative)

Kedua-dua kaedah pengelompokan memerlukan pembangunan aturcara dilakukan dengan menggunakan bahasa pengaturcaraan C.

3.5.4 Peramalan Taburan Hujan

Prosidur ini adalah bertujuan untuk melakukan pengujian ke atas kelompok atau kumpulan data kajicuaca yang telah dikelompokkan dengan menggunakan teknik peraturan kesatuan dan juga kaedah statistik. Pengujian dilaksanakan dengan menggunakan pakej NeuNetPro versi 2.3, di mana peramalan taburan hujan akan dilakukan di dalam proses ini. Keputusan peramalan taburan hujan akan menunjukkan keberkesanan dan juga ketepatan operasi pengelompokan yang dilakukan dengan menggunakan teknik peraturan kesatuan dan kaedah statistik.

3.5.5 Penganalisaan dan Penilaian

Dalam fasa penganalisaan dan penilaian ini pula, hasil peramalan taburan hujan yang diperolehi akan dianalisa. Keputusan peramalan yang dihasilkan daripada perlaksanaan peramalan terhadap pengelompokan data kajicuaca dengan menggunakan teknik peraturan kesatuan dan kaedah statistik akan dinilai. Ianya bertujuan untuk melihat sejauhmanakah ketepatan dan keberkesanan di antara teknik peraturan kesatuan dengan kaedah statistik di dalam mengelompokkan data kajicuaca.

3.5.6 Perbandingan

Pengukuran ketepatan dan keberkesanan hasil peramalan ini dilakukan berdasarkan kepada perbandingan nilai ralat RMS dan pekali korelasi yang diperolehi di dalam setiap eksperimen yang dijalankan. Perbandingan ini dilaksanakan untuk melihat sejauhmanakah keberkesanan dan ketepatan peramalan taburan hujan dengan menggunakan data kajicuaca yang dikelompokkan menggunakan teknik peraturan kesatuan dan kaedah statistik. Ianya bertujuan untuk mengetahui kaedah

pengelompokan manakah yang menghasilkan peramalan taburan hujan yang lebih tepat. Akhir sekali, graf yang menunjukkan nilai RMS dan pekali korelasi yang dihasilkan dalam setiap eksperimen yang telah dijalankan menggunakan teknik peraturan kesatuan dan juga kaedah statistik bagi tujuan pengelompokkan akan diplotkan untuk melihat perbezaan ketepatan hasil peramalan.

3.6 Kesimpulan

Kajian ini adalah bertujuan untuk melihat sejauhmanakah keberkesanan di antara teknik pengelompokan partitional (algoritma Forgy dan K-Means) dan pengelompokan hierarchical (algoritma agglomerative) ini di dalam melakukan pengelompokan data-data kajicuaca. Bagi tujuan perbandingan ketepatan di antara dua teknik ini, pakej NeuNet Pro digunakan untuk melakukan peramalan terhadap data-data taburan hujan berdasarkan kepada kelompok-kelompok yang telah dihasilkan melalui pembangunan aturcara C. Diharapkan kajian ini dapat memberikan sedikit sumbangan kepada pihak Jabatan Perkhidmatan Kajicuaca Malaysia di dalam membantu melakukan kerja-kerja penganalisaan data bagi tujuan peramalan taburan hujan dan juga pemantauan perubahan cuaca.

3.7 Ringkasan

Bab ini menerangkan metodologi yang akan digunakan di dalam menjalankan kajian ini. Metodologi merupakan gabungan kaedah, polisi, prosidur, peraturan, piawai, teknik, bahasa pengaturcaraan dan metodologi-metodologi lain yang digunakan untuk menganalisa dan memperincikan keperluan kajian. Sumber data dan peralatan yang dipilih untuk kajian ini juga ditentukan melalui pemerhatian dan juga pembacaan artikel-artikel yang berkaitan.

Selain daripada itu, prosidur, analisa dan pengujian yang digunakan di dalam kajian ini juga dikaji dan enam fasa telah dikenalpasti bagi tujuan melakukan kajian ini. Di antaranya ialah analisa keperluan, perolehan dan pemprosesan data, pengelompokan data kajicuaca, peramalan taburan hujan, penganalisaan serta penilaian dan akhir sekali ialah perbandingan hasil peramalan.

BAB IV

PENGELOMPOKAN MENGGUNAKAN PARTITIONAL (K-MEANS) DAN HIERARCHICAL (AGGLOMERATIVE)

4.1 Pengenalan

Teknik K-Means dan Agglomerative merupakan suatu pendekatan biasanya digunakan untuk mendapatkan maklumat yang boleh dipercayai dalam membuat sebarang peramalan atau keputusan dalam pelbagai bidang. Dewasa ini, saiz pangkalan data yang mengandungi pelbagai jenis atribut data semakin meningkat. Oleh yang demikian, atribut-atribut ini perlu dikelompokkan kepada koleksi atribut-atribut yang mempunyai persamaan. Ianya bertujuan untuk mendapatkan satu ringkasan bagi pangkalan data tersebut atau sebagai persediaan untuk melakukan operasi seterusnya, contohnya seperti peramalan.

Pengelompokan ialah membahagikan satu pangkalan data kepada beberapa kelompok yang mempunyai ciri atau sifat yang sama. Apabila atribut data yang mempunyai ciri yang sama telah dikumpulkan ke dalam satu kelompok, ini bermakna atribut data yang berada di dalam kelompok yang lain mempunyai ciri yang berbeza daripada mereka. Atribut-atribut di dalam satu kelompok mestilah mempunyai persamaan yang boleh diukur menggunakan kaedah tertentu, contohnya mengira jarak di antara atribut-atribut tersebut.

4.2 Pengelompokkan Partitional

Bahagian ini akan menerangkan mengenai pembangunan aturcara algoritma K-Means, eksperimen dan juga perbincangan.

4.2.1 Pembangunan Algoritma K-Means

Set data kaji cuaca mempunyai pelbagai data numerik yang diambil daripada pengukuran menggunakan peralatan kaji cuaca (Malaysia Meteorology Services, 2004). Di dalam eksperimen ini, 500 data pemerhatian kaji cuaca diambil daripada PKM bagi tujuan pengujian bermula daripada 1 September 2000 sehingga 21 September 2000. Data-data diambil mengikut jam dan terdapat tujuh pembolehubah yang dikenalpasti iaitu *windvane*, *humidity*, *energy*, *temp*, *tension*, *radiation* dan *windspeed*. Jadual 4.1 menunjukkan contoh sampel data kaji cuaca yang digunakan di dalam analisa.

Jadual 4.1: Sampel Data Kaji Cuaca

ID	Date	Time	Windvane	Humidity	……	Windspeed
1	8/1/00	0:00	197.0	0.0		0.0
2	8/1/00	1:00	201.2	0.0		0.0
3	8/1/00	2:00	206.5	0.0		0.0
…	…	…	…	…		…
…	…	…	…	…		…
500	8/21/00	19:00	290.9	1.5		0.0

Di dalam proses pembangunan algoritma K-Means, penggunaan bahasa pengaturcaraan C digunakan bagi tujuan pengujian ke atas data. Rajah 4.1, 4.2 dan 4.3 menunjukkan keratan aturcara bagi algoritma K-Means.

```
void initkmeans()
    {for(int j=0;j<k;j++)
            {clusterdata[j].member[0]=j;
             clusterdata[j].count=1;
             for(int i=0;i<jum;i++)
                 {clusterdata[j].centroid[i]=itemdata[j].item[i];
                  }
        }
    }
```

Rajah 4.1: Fungsi initkmeans( )

Berdasarkan kepada fungsi initkmeans( ), ia akan mengumpukkan k kelompok daripada input pengguna kepada k kelompok yang pertama daripada data kaji cuaca. Fungsi ini akan menilaiawalkan k kelompok sebagai kelompok asas kepada struktur clusterdata[ ] seperti yang dinyatakan pada keratin aturcara ini; clusterdata[j].centroid[i]=itemdata[j].item[i].

```
void assigncluster()
        {
        float min_dist;
        int index;
        for(int i=k;i<7;i++) //sample
            {
            min_dist=euclidist(i,0);
            index=0;
            //closest mean
            for(int j=1;j<k;j++)   //cluster
                    {
                    clusterdata[j].dist=euclidist(i,j);
                    if(clusterdata[j].dist<min_dist || clusterdata[j].dist==0)
                            {
                                    min_dist=clusterdata[j].dist;
                                    index=j;
                                    }
                    }
             //add member
            clusterdata[index].count=clusterdata[index].count+1;
            clusterdata[index].member[clusterdata[index].count-1]=i;
            //mean
             float ctd=0;
            for(int x=0;x<jum;x++)
                    {
                    ctd=0;
                     for(int y=0;y<clusterdata[index].count;y++)
                            {

ctd=ctd+itemdata[clusterdata[index].member[y]].item[x];
                                    }
                     clusterdata[index].centroid[x]=ctd/clusterdata[index].count;
                            ctd=0;
                    }
            }
        }
```

Rajah 4.2: Fungsi assigncluster( )

Assigncluster( ) akan mengumpukkan setiap data pembolehubah kepada k kelompok melalui fungsi initkmeans( ). Ia akan mengumpukkan semua pembolehubah kepada kelompok yang sama menggunakan matriks jarak Euclidean di mana persamaan data dengan kelompok akan dikira. Centroid atau min juga akan dikira.

```
void checkcluster()
        {int ctr;
        iter_kmeans=0;
        do
                {iter_kmeans=iter_kmeans+1;
                 for(int c=0;c<7;c++)
                        {checkingdata[c]=cl[c];
                        }
                checkingcluster();
                ctr=0;
                for(int x=0;x<k;x++)
                        {for(int y=0;y<7;y++)
                                {if(checkingdata[x].member[y]!=clusterdata[x].member[y])
                                        {ctr=1;
                                        }
                                }
                        }
                if(ctr==1)
                        {for(int c=0;c<7;c++)
                                {clusterdata[c]=checkingdata[c];
                                }
                        }
                }while(ctr==1);
         for(int c=0;c<7;c++)
                {clusterdata[c]=checkingdata[c];
                }
        }
void checkingcluster()
        {float min_dist;
        int index;
        for(int i=0;i<7;i++)
                {index=0;
                min_dist=euclidist_check(i,0);
                for(int j=1;j<k;j++)
                        {checkingdata[j].dist=euclidist_check(i,j);
                        if(checkingdata[j].dist<min_dist || checkingdata[j].dist==0)
                                {min_dist=checkingdata[j].dist;
                                index=j;
                                }
                        }
                 //add member
                checkingdata[index].count=checkingdata[index].count+1;
                checkingdata[index].member[checkingdata[index].count-1]=i;
                 //mean
                 float ctd=0;
                for(int x=0;x<jum;x++)
                        {ctd=0;
                        for(int y=0;y<checkingdata[index].count;y++)
                                {ctd=ctd+itemdata[checkingdata[index].member[y]].item[x];
                                 }
                        checkingdata[index].centroid[x]=ctd/checkingdata[index].count;
                        ctd=0;
                        }
                }
        }
```

Rajah 4.3: Fungsi checkcluster( ) dan checkingcluster( )

Di dalam fungsi checkcluster( ), ia akan menyemak kembali semua pembolehubah dengan kelompok yang wujud dengan mengira persamaan di antara pembolehubah dan kelompok. Proses ini akan menyemak ahli yang terdapat di dalam kelompok sebelum dan selepas fungsi checkingcluster( ) di mana ia akan mengira dan mengumpukkan data kepada kelompok yang terdekat. Proses ini akan berterusan sehingga semua ahli di dalam kelompok adalah sama sebelum dan selepas fungsi checkingcluster( ). Rajah 5.4 menunjukkan contoh output yang terhasil dengan 4 kelompok:

```
~~~~~~~~~~~~~~~~~~
K-MEANS METHOD
~~~~~~~~~~~~~~~~~~

Number of clusters : 4

Result K-Means

C1 - [ windvane ]
C2 - [ tension ]
C3 - [ energy  radiation ]
C4 - [ humidity  temp  windspeed ]
Iteration time of K-means : 2
```

Rajah 5.4: Output aturcara

Rajah 5.4 menunjukkan output apabila pengguna memasukkan input 4 kelompok. Berdasarkan output tersebut, dapat dilihat di mana daripada 7 pembolehubah telah dikategorikan kepada 4 kelompok yang sama.

4.2.2 Eksperimen

Kajian ini melibatkan penggunaan salah satu teknik pengelompokan data, iaitu algoritma K-Means untuk menghasilkan set-set kelompok data kaji cuaca. Penentuan set kelompok data kaji cuaca ini dilakukan berdasarkan kepada pengiraan jarak Euclidean yang telah diaplikasikan dalam aturcara yang telah dibangunkan. Sehubungan dengan itu, proses pembahagian atau penentuan set kelompok ini dilakukan sebanyak lima(5)

kali, di mana ia melibatkan 100, 200, 300, 400 dan 500 set data kaji cuaca. Jadual 4.2 di bawah menunjukkan sampel data kaji cuaca yang digunakan di dalam eksperimen ini.

Jadual 4.2: Sampel Data Kaji Cuaca

ID	Tarikh	Masa	Windvane	Humidity	Energy	Temp	Tension	Radiation	Windspeed	Rainfall
1	8/1/00	0:00	197	0	-0.58	-10.8	-4.8	0	0	0
2	8/1/00	1:00	201.2	0	-0.63	-11	-4.8	-48.828	0	0
3	8/1/00	2:00	206.5	0	-0.63	-11.1	-5.8	-48.828	0	0
4	8/1/00	3:00	235.5	0	-0.24	-11.4	-5.8	-48.828	0	0
5	8/1/00	4:00	293.7	0	-0.39	-10.9	-4.8	-48.828	0	0
6	8/1/00	5:00	143.2	0	-0.14	-7.8	-4.8	-48.828	0	0
7	8/1/00	6:00	201.2	0	0.04	-6.7	-3.9	-48.828	0	0
8	8/1/00	7:00	343.8	0	0.53	-6.8	0.9	-48.828	0	2.5
9	8/1/00	8:00	261.2	0	42.23	-7.6	0.9	0	0	0.5
10	8/1/00	9:00	304.7	0	58.88	-9.1	-2.9	97.656	0	0
…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…
500	8/21/00	19:00	290.0	1.5	0.04	18.2	0	0	0	0

Bagi setiap kategori (100, 200, 300, 400 dan 500) set data kaji cuaca tersebut, proses pengelompokan ini dilakukan berulangkali bagi menghasilkan kelompok 2, kelompok 3, kelompok 4, kelompok 5 dan kelompok 6. Hasil penentuan kelompok-kelompok ini boleh dirujuk pada **Jadual 4.3** berikut.

Jadual 4.3: Penentuan Kelompok Data Kaji Cuac**a**

Bil. Kelompok	**100 set**	**200 set**	**300 set**	**400 set**	**500 set**
2	(a) (b, c, d, e, f, g)	(a) (b, c, d, e, f, g)	(a) (b, c, d, e, f, g)	(a) (b, c, d, e, f, g)	(a) (b, c, d, e, f, g)
3	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)
4	(a) (e) (c, f) (b, d, g)	(a) (e) (c, f) (b, d, g)	(a) (e) (c, f) (b, d, g)	(a) (e) (c, f) (b, d, g)	(a) (b, e, g) (c, f) (d)
5	(a) (b, g) (c, f) (d) (e)	(a) (b, g) (c, f) (d) (e)	(a) (b, g) (c, f) (d) (e)	(a) (b) (c, f) (d, g) (e)	(a) (b) (c, f) (d, g) (e)
6	(a) (b, g) (c) (d) (e) (f)	(a) (b, g) (c) (d) (e) (f)	(a) (b, g) (c) (d) (e) (f)	(a) (b) (c) (d, g) (e) (f)	(a) (b) (c) (d, g) (e) (f)

Di mana;

a – *windvane*, b – *humidity*, c – *energy*, d – *temp*, e – *tension*, f – *radiation*, g - *windspeed*

Setelah hasil penentuan kelompok-kelompok data kaji cuaca telah dilakukan, pengujian terhadap kelompok-kelompok tersebut pula dilaksanakan. Ini bertujuan untuk melihat keberkesanan algoritma K-Means di dalam mengelompokkan data kaji cuaca. Pengujian terhadap hasil pengelompokan ini dilakukan dengan menggunakan atribut-atribut data kaji cuaca yang berada di dalam kelompok yang berlainan sebagai data input untuk melakukan proses peramalan taburan hujan. Proses peramalan taburan hujan ini dilakukan dengan menggunakan pakej perisian NeuNetPro.

Oleh yang demikian, beberapa eksperimen peramalan taburan hujan telah dilaksanakan, di mana ia bertujuan untuk melihat ketepatan hasil peramalan taburan hujan tersebut. Perlaksanaan eksperimen ini dilakukan dengan menggunakan jumlah set data kaji cuaca yang sama bagi bilangan kelompok yang berbeza untuk meramal taburan hujan. Untuk tujuan tersebut, terdapat lima(5) set eksperimen telah dikenalpasti dan dilaksanakan, di antaranya;

i) Eksperimen Pertama – melibatkan 100 set data kaji cuaca bagi kelompok 2, 3, 4, 5 dan 6.
ii) Eksperimen Kedua – melibatkan 200 set data kaji cuaca bagi kelompok 2, 3, 4, 5 dan 6.
iii) Eksperimen Ketiga - melibatkan 300 set data kaji cuaca bagi kelompok 2, 3, 4, 5 dan 6.
iv) Eksperimen Keempat - melibatkan 400 set data kaji cuaca bagi kelompok 2, 3, 4, 5 dan 6.
v) Eksperimen Kelima - melibatkan 500 set data kaji cuaca bagi kelompok 2, 3, 4, 5 dan 6.

Bagi eksperimen yang melibatkan kelompok 2, penulis telah melaksanakan enam eksperimen yang berasingan di mana ia melibatkan penggunaan atribut (a,b), atribut (a,c), atribut (a,d), atribut (a,e), atribut (a,f) dan atribut (a,g) sebagai data input untuk melakukan proses peramalan taburan hujan. Berdasarkan nilai RMS dan pekali korelasi yang dihasilkan di dalam keenam-enam eskperimen tersebut, purata bagi nilai RMS dan pekali korelasi dikira dan ambil sebagai nilai RMS dan pekali korelasi bagi kelompok 2.

Manakala bagi eksperimen yang melibatkan kelompok 3, penulis telah melaksanakan lapan eksperimen berasingan yang melibatkan atribut (a,b,c), atribut (a,d,c), atribut (a,e,c), atribut (a,g,c), atribut (a,b,f), atribut (a,d,f), atribut (a,e,f) dan atribut (a,g,f) sebagai data input untuk melakukan peramalan. Kemudian, purata nilai RMS dan pekali korelasi dikira bagi kesemua eksperimen tersebut dan diambil sebagai nilai RMS dan pekali korelasi bagi kelompok 3.

Seterusnya, enam eksperimen bagi kelompok 4 pula dilaksanakan di mana ia melibatkan atribut (a,e,c,b), atribut (a,e,c,d), atribut (a,e,c,g), atribut (a, e,f,b), atribut (a,e,f,d) dan atribut (a,e,f,g). Purata nilai RMS dan pekali korelasinya dikira dan diambil sebagai nilai RMS dan pekali korelasi bagi kelompok 4. Empat eksperimen berikutnya pula dijalankan di mana ia melibatkan kelompok 5. Oleh yang demikian, atribut (a,b,c,d,e), atribut (a,g,c,d,e), atribut (a,b,f,d,e) dan atribut (a,g,f,d,e) telah digunakan sebagai data input kepada proses peramalan taburan hujan. Dan purata nilai RMS dan pekali korelasinya telah diambil sebagai nilai RMS dan pekali korelasi bagi kelompok 5.

Dan akhir sekali, dua eksperimen bagi kelompok 6 telah dilakukan di mana ia melibatkan atribut (a,b,c,d,e,f) dan atribut (a,g,c,d,e,f) sebagai input kepada proses peramalan taburan hujan. Kemudian, purata nilai RMS dan pekali korelasinya telah diambil sebagai nilai RMS dan pekali korelasi bagi kelompok 6. Jadual 4.4(a), 4.4(b), 4.4(c), 4.4(d) dan 4.4(e) pada **Lampiran A** menunjukkan keputusan peramalan taburan hujan yang dihasilkan di dalam Eksperimen Pertama hingga Eksperimen Kelima di atas.

4.2.3 Perbincangan

Di dalam kajian ini, eksperimen peramalan taburan hujan yang telah dijalankan adalah bertujuan untuk melihat keupayaan algoritma K-Means di dalam memberikan nilai ramalan yang tepat ke atas data taburan hujan. Berdasarkan keputusan peramalan yang telah dihasilkan tersebut, didapati keputusan peramalan taburan hujan yang dihasilkan menunjukkan semakin besar bilangan kelompok data kaji cuaca yang digunakan sebagai data input kepada proses peramalan, semakin tinggi prestasi peramalan taburan hujan

yang dihasilkan, bagi semua set data (jumlah berlainan) yang digunakan. Ini dibuktikan oleh nilai RMS yang semakin berkurangan dan juga nilai pekali korelasi yang semakin tinggi (bagi kelompok 2, 3, 4, 5, dan 6). Selain daripada itu, hasil peramalan juga menunjukkan semakin banyak jumlah data set kaji cuaca yang digunakan untuk melakukan peramalan taburan hujan, semakin tinggi prestasi peramalan yang dihasilkan. Prestasi peramalan ini ditunjukkan oleh keputusan nilai RMSnya yang semakin berkurangan, manakala nilai pekali korelasinya yang semakin meningkat bagi semua kelompok 2, 3, 4, 5 dan 6.

Walau bagaimanapun, terdapat beberapa masalah yang timbul semasa perlaksanan eksperimen ini, di antaranya ialah;

i) Sebahagian daripada data-data kaji cuaca yang digunakan di dalam kajian ini adalah data melampau (data melampau bermaksud julat antara data yang berturutan adalah bersaiz besar). Oleh yang demikian, ini telah menyebabkan hasil peramalan taburan hujan kurang memuaskan.
ii) Masalah kelemahan pakej NeuNetPro yang digunakan untuk pengujian peramalan taburan hujan. Ini kerana pakej tersebut tidak dapat digunakan untuk membuat peramalan masa hadapan. Selain daripada itu, pakej ini juga tidak dapat menjanakan proses peramalan jika bilangan data input yang digunakan, kurang daripada 10 data.

Di samping itu, penulis juga telah membuat andaian bagi menjayakan eksperimen yang telah dijalankan di dalam kajian ini. Di antaranya ialah data-data kaji cuaca yang digunakan di dalam eksperimen ini dianggap bersih dan bebas dari hingar. Selain daripada itu, jarak di antara atribut-atribut data kaji cuaca yang paling kecil dianggap mempunyai ciri-ciri persamaan yang kuat dan sebaliknya.

4.3 Pengelompokkan Hiearchical

Bahagian ini akan menerangkan mengenai pembangunan metodologi di dalam pengelompokan hierarchical dan juga eksperimen yang dilakukan.

4.3.1 Metodologi

Perlaksanaan kajian ini telah dilakukan mengikut beberapa aktiviti. Di antara aktiviti-aktiviti yang terlibat ialah mengumpul dan menganalisis data kaji cuaca, mengira jarak Euclidean, membangunkan atur cara, mengelompokkan set data kaji cuaca, melakukan pengujian kelompok data kaji cuaca, menganalisis keputusan pengujian dan akhir sekali ialah membuat perbandingan keputusan pengujian yang telah dihasilkan.

Bagi menjayakan perlaksanaan kajian ini, sebanyak 500 set data kaji cuaca, bermula dari 1 September 2000 sehingga 21 September 2000 telah dikumpul. Data tersebut telah diperolehi daripada Jabatan Perkhidmatan Kaji cuaca Malaysia. Terdapat lapan pemboleh ubah atau atribut utama data kaji cuaca yang telah diguna pakai di dalam perlaksanaan kajian ini, di antaranya ialah *windvane*, *humidity*, *energy*, *temp*, *tension*, *radiation*, *windspeed* dan *rainfall*.

Selepas data kaji cuaca tersebut dianalisis, pengiraan jarak Euclidean pula dilakukan. Proses ini dijalankan untuk mendapatkan persamaan di antara atribut-atribut data kaji cuaca tersebut untuk tujuan mengelompokkan atribut-atribut yang berkaitan. Formula yang digunakan untuk melakukan pengiraan jarak *Euclidean* ialah :

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \qquad \ldots(4)$$

di mana $x = (x_1, x_2, ..., x_n)$ dan $y = (y_1, y_2, ..., y_n)$

Berdasarkan pengiraan jarak Euclidean yang telah dihasilkan, algoritma pengelompokan hierarki bagi *single link*, *average link* dan *complete link* dibangunkan

menggunakan bahasa pengaturcaraan C. Dengan menggunakan atur cara yang telah dibangunkan, atribut-atribut data kaji cuaca (kecuali atribut *rainfall*) dikelompokkan kepada beberapa kelompok iaitu kelompok 2, 3, 4, 5 dan 6. Proses ini dilakukan berulang-ulang kali dengan menggunakan beberapa set data yang terdiri daripada 100 set data, 200 set data, 300 set data, 400 set data dan 500 set data kaji cuaca.

Setelah pengelompokan dijalankan ke atas set data kaji cuaca, setiap kelompok yang terhasil diuji dengan menggunakan Pakej NeuNet Pro 2.3. Pengujian ini telah dibahagikan kepada dua kategori, iaitu eksperimen yang melibatkan atribut-atribut dari kelompok yang sama dan juga eksperimen yang melibatkan atribut-atribut dari kelompok yang berlainan. Di dalam pengujian ini, atribut-atribut dari kelompok yang sama dan atribut-atribut dari kelompok yang berlainan digunakan sebagai data input untuk melakukan proses peramalan taburan hujan. Pengujian kelompok ini dilakukan untuk membuat penganalisisan dan perbandingan prestasi peramalan taburan hujan di antara set-set eksperimen yang dijalankan.

Penganalisisan kemudian dilakukan ke atas hasil pengujian tersebut untuk menentukan prestasi teknik pengelompokan menggunakan algoritma pengelompokan hierarki. Penganalisaan ini dilakukan dengan melihat nilai perbezaan di antara nilai sebenar taburan hujan serta nilai ramalan taburan hujan yang dihasilkan. Berdasarkan kepada perbezaan nilai tersebut, nilai ralat min punca kuasa dua (RMS) dan pekali korelasi akan dihasilkan. Sehubungan dengan itu, nilai RMS dan pekali korelasi ini digunakan sebagai perbandingan bagi menentukan keberkesanan teknik pengelompokan hierarki.

4.3.2 Eksperimen

Di dalam kajian yang dijalankan, proses pengelompokan menggunakan algoritma *Agglomerative* telah dilakukan ke atas 500 set data kaji cuaca. Proses pengelompokan ini telah dibahagikan kepada tiga bahagian utama iaitu pengelompokan menggunakan

algoritma *Single Link*, algoritma *Average Link* dan algoritma *Complete Link*. Eksperimen ini dijalankan berulang-ulang kali dengan mengelompokkan 100 set data, 200 set data, 300 set data, 400 set data dan 500 set data kaji cuaca kepada 2 kelompok, 3 kelompok, 4 kelompok, 5 kelompok dan 6 kelompok.

Selepas pengelompokan dijalankan ke atas set-set data tadi, pengujian kelompok pula dilakukan bagi setiap kelompok data kaji cuaca yang telah dihasilkan. Ia melibatkan pengujian peramalan taburan hujan dengan menggunakan atribut-atribut dari kelompok data kaji cuaca yang sama dan juga kelompok data kaji cuaca yang berlainan. Pengujian ini dilakukan dengan menggunakan aplikasi NeuNet Pro 2.3. Keputusan peramalan taburan hujan yang diperolehi digunakan untuk membuat penganalisisan dan perbandingan.

4.3.2.1 Algoritma *Single Link*

Berdasarkan eksperimen yang telah dilakukan, Jadual 4.5 berikut adalah merupakan hasil pengelompokan data kaji cuaca dengan menggunakan algoritma *Single Link*.

Jadual 4.5 : Hasil pengelompokan data kaji cuaca menggunakan algoritma *Single Link*

Bil. Kelompok	100 Set Data	200 Set Data	300 Set Data	400 Set Data	500 Set Data
2	(a) (b, c,d, e, f, g)	(a, b, d, e, g) (c, f)	(a, b, d, e, g) (c, f)	(a, b, d, e, g) (c, f)	(a, b, d, e, g) (c, f)
3	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)
4	(a) (b, d, e, g) (c) (f)	(a) (b, d, e, g) (c) (f)	(a) (b, d, e, g) (c) (f)	(a) (b, d, e, g) (c) (f)	(a) (b, d, e, g) (c) (f)
5	(a) (b, d, g) (c) (e) (f)	(a) (b, d, g) (c) (e) (f)	(a) (b, d, g) (c) (e) (f)	(a) (b, d, g) (c) (e) (f)	(a) (b, d, g) (c) (e) (f)

6	(a) (b, g) (c) (d) (e) (f)	(a) (b, g) (c) (d) (e) (f)	(a) (b, g) (c) (d) (e) (f)	(a) (b) (c) (d, g) (e) (f)	(a) (b) (c) (d, g) (e) (f)

a – *windvane*, b – *humidity*, c – *energy*, d – *temp*, e – *tension*, f – *radiation*, g – *windspeed*

Setelah kelompok-kelompok data kaji cuaca dikenalpasti, pengujian kelompok pula dilakukan. Proses pengujian ini melibatkan penggunaan atribut-atribut data kaji cuaca (kecuali atribut *rainfall*) sebagai data input untuk melakukan proses peramalan taburan hujan. Pengujian ini telah dibahagikan kepada dua kategori, iaitu pengujian bagi kelompok data kaji cuaca yang sama dan juga pengujian bagi kelompok data kaji cuaca yang berlainan. Oleh yang demikian, atribut-atribut dari kelompok yang sama dan juga atribut-atribut dari kelompok yang berlainan telah digunakan sebagai data input kepada proses peramalan tersebut. Bagi setiap kategori pengujian, terdapat beberapa eksperimen yang telah dijalankan.

Setiap eksperimen tersebut melibatkan atribut-atribut bagi 2 kelompok, 3 kelompok, 4 kelompok, 5 kelompok dan 6 kelompok data kaji cuaca dan juga menggunakan 100 set data, 200 set data, 300 set data, 400 set data dan 500 set data kaji cuaca. Pengujian bagi setiap kategori kelompok dan set data dilakukan beberapa kali dan nilai purata bagi nilai ralat min punca kuasa dua (RMS) dan pekali korelasi telah diambil dan digunakan untuk tujuan perbandingan. Jadual 4.6(a) dan 4.6(b) menunjukkan hasil peramalan taburan hujan yang dilakukan ke atas kelompok atribut data kaji cuaca yang sama dan juga kelompok atribut data kaji cuaca yang berlainan.

Jadual 4.6(a): Keputusan peramalan taburan hujan bagi kelompok berlainan

KELOMPOK	**100 Set Data**		**200 Set Data**		**300 Set Data**		**400 Set Data**		**500 Set Data**	
	RMS	CC	RMS	CC	RMS	CC	RMS	CC	RMS	CC
2 Kelompok	0.95	0.07	0.85	0.27	0.88	0.17	0.96	0.12	0.95	0.09
3 Kelompok	0.93	0.15	0.81	0.39	0.85	0.26	0.96	0.16	0.94	0.16
4 Kelompok	0.88	0.37	0.79	0.46	0.85	0.23	0.92	0.33	0.92	0.28
5 Kelompok	0.91	0.22	0.78	0.49	0.78	0.49	0.89	0.25	0.90	0.34
6 Kelompok	0.84	0.38	0.68	0.64	0.74	0.56	0.79	0.67	0.78	0.56

Berdasarkan hasil pengujian bagi kelompok data kaji cuaca berlainan yang telah dilaksanakan, maka penganalisisan terhadap keputusan peramalan taburan hujan telah dibuat. Secara keseluruhannya, didapati kesemua eksperimen yang melibatkan set data 100, 200, 300, 400 dan 500 menunjukkan bahawa hasil atau prestasi peramalan semakin meningkat, selari dengan bilangan atribut-atribut data kaji cuaca yang digunakan di dalam pengujian tersebut. Atau dengan kata lain, semakin banyak bilangan atribut data kaji cuaca yang digunakan sebagai data input kepada proses peramalan taburan hujan tersebut, semakin baik prestasi peramalan yang dihasilkan. Ini dibuktikan oleh nilai RMS dan juga pekali korelasi yang telah dihasilkan di dalam perlaksanaan eksperimen yang berkaitan.

Selain daripada itu, hasil eksperimen juga menunjukkan keputusan yang dihasilkan oleh eksperimen yang melibatkan set data kaji cuaca 200 memberikan keputusan peramalan yang terbaik berbanding set data kaji cuaca yang lain bagi semua kelompok 2, 3, 4, 5 dan 6 yang dijalankan. Keadaan ini mungkin juga disebabkan oleh kewujudan bilangan data melampau yang banyak di dalam set data 300, 400 dan 500.

Jadual 4.6(b): Keputusan peramalan taburan hujan bagi kelompok sama

KELOMPOK	**100 Set Data**		**200 Set Data**		**300 Set Data**		**400 Set Data**		**500 Set Data**	
	RMS	CC	RMS	CC	RMS	CC	RMS	CC	RMS	CC
2 Kelompok	0.87	0.29	0.82	0.29	0.77	0.36	0.88	0.30	0.90	0.28
3 Kelompok	0.94	0.12	0.89	0.08	0.84	0.19	0.85	0.21	0.95	0.13
4 Kelompok	0.70	0.07	0.67	0.06	0.63	0.14	0.77	0.16	0.69	0.14
5 Kelompok	0.95	0.03	0.86	0.15	0.86	0.14	1.03	0.07	0.96	0.08
6 Kelompok	0.95	0.03	0.90	0.06	0.88	0.10	0.93	0.03	0.80	0.06

Manakala bagi eksperimen yang melibatkan penggunaan atribut-atribut dari kelompok data kaji cuaca yang sama pula menunjukkan keputusan peramalan taburan hujan yang berbeza. Berdasarkan kepada keputusan pada Jadual 2(b) di atas, maka dapatlah dirumuskan bahawa keputusan peramalan yang dihasilkan oleh 4 kelompok memberikan prestasi yang terbaik berbanding dengan empat eksperimen yang lain. Prestasi ini boleh dilihat pada nilai RMS dan juga nilai pekali korelasi yang dihasilkan oleh eksperimen yang melibatkan penggunaaan atribut-atribut data kaji cuaca dari 4 kelompok di dalam kesemua eskperimen sama ada yang melibatkan 100 data set, 200

data set, 300 data set, 400 data set, mahupun 500 data set. Di samping itu, keputusan peramalan taburan hujan yang ditunjukkan oleh eksperimen yang melibatkan 300 set data kaji cuaca juga memberikan keputusan prestasi peramalan yang lebih baik jika dibandingkan dengan prestasi peramalan taburan hujan yang melibatkan set data yang lain.

Oleh yang demikian, secara keseluruhannya bolehlah dirumuskan bahawa keputusan peramalan taburan hujan yang dihasilkan oleh eskperimen yang melibatkan penggunaan atribut dari kelompok data kaji cuaca yang berlainan menunjukkan prestasi peramalan taburan hujan yang lebih baik berbanding dengan prestasi peramalan taburan hujan yang dihasilkan oleh eksperimen yang melibatkan penggunaan atribut dari kelompok data kaji cuaca yang sama. Ini dibuktikan oleh keputusan peramalan taburan hujan yang dihasilkan di dalam kelima-lima eksperimen yang melibatkan bilangan set data kaji cuaca yang berbeza.

Selain daripada itu, keputusan peramalan bagi eksperimen kelompok berlainan menunjukkan semakin banyak bilangan atribut data kaji cuaca yang digunakan di dalam pengujian tersebut, semakin tinggi prestasi peramalan yang dihasilkan. Ini dapat dilihat pada nilai RMS yang semakin berkurang dan juga nilai pekali korelasi yang semakin meningkat.

4.3.2.2 Algoritma *Average Link*

Proses pengelompokan telah dilaksanakan ke atas set data kaji cuaca dengan menggunakan algoritma *Average Link*. Jadual 4.7 menunjukkan hasil pengelompokan data kaji cuaca tersebut.

Jadual 4.7: Hasil pengelompokan data kaji cuaca menggunakan algoritma *Average Link*

Bil. Kelompok	100 Set Data	200 Set Data	300 Set Data	400 Set Data	500 Set Data
2	(a,c,d,e,f,g) (b)	(a,c,d,e,f,g) (b)	(a,c,d,e,f,g) (b)	(a,c,d,e,f,g) (b)	(a,c,d,e,f,g) (b)
3	(a) (b) (c,d,e,f,g)	(a) (b,c, d, e, f) (g)	(a, c, d, e, g) (b) (f)	(a, c, d, e, g) (b) (f)	(a, c, d, e, g) (b) (f)
4	(a) (b) (c,d,f,g) (e)	(a) (b,c,d, f) (e) (g)	(a, c, d, e) (b) (f) (g)	(a, c, d, e) (b) (f) (g)	(a, c, d, e) (b) (f) (g)
5	(a) (b) (c,f,g) (d) (e)	(a) (b) (c, d , f) (e) (g)	(a, e, d) (b) (e) (f) (g)	(a, e, d) (b) (e) (f) (g)	(a, e, d) (b) (e) (f) (g)
6	(a) (b) (c,g) (d) (e) (f)	(a) (b) (c, d) (e) (f) (g)	(a) (b) (c, d) (e) (f) (g)	(a) (b) (c, d) (e) (f) (g)	(a) (b) (c, d) (e) (f) (g)

a – *windvane*, b – *humidity*, c – *energy*, d – *temp*, e – *tension*, f – *radiation*, g – *windspeed*

Berdasarkan kelompok-kelompok data kaji cuaca yang telah dihasilkan dengan menggunakan algoritma *Average Link*, pengujian kelompok dilaksanakan dengan menjalankan beberapa eksperimen. Di antaranya melibatkan penggunaan atribut dari kelompok yang sama dan juga dari kelompok yang berlainan. Di dalam perlaksanaan eksperimen tersebut, atribut data kaji cuaca (atribut selain *rainfall*) telah digunakan sebagai data input kepada proses pengujian, iaitu proses peramalan taburan hujan. Keputusan yang dihasilkan di dalam setiap eskperimen tersebut, akan dicatatkan dan nilai purata bagi nilai ralat min punca kuasa dua (RMS) dan juga nilai pekali korelasi digunakan untuk perbandingan keputusan. Jadual 4.8(a) dan 4.8(b) menunjukkan hasil purata bagi nilai RMS dan pekali korelasi yang dihasilkan oleh setiap eksperimen yang dijalankan.

Jadual 4.8(a): Keputusan peramalan taburan hujan bagi kelompok berlainan

KELOMPOK	100 Set Data		200 Set Data		300 Set Data		400 Set Data		500 Set Data	
	RMS	CC	RMS	CC	RMS	CC	RMS	CC	RMS	CC
2 Kelompok	0.94	0.09	0.87	0.24	0.74	0.15	1.09	0.14	0.96	0.10
3 Kelompok	0.94	0.13	0.83	0.35	0.86	0.23	0.94	0.26	0.78	0.15
4 Kelompok	0.95	0.09	0.85	0.30	0.74	0.54	0.86	0.62	0.81	0.53
5 Kelompok	0.89	0.33	0.70	0.62	0.68	0.64	0.77	0.71	0.74	0.63
6 Kelompok	0.94	0.17	0.58	0.76	0.75	0.55	0.76	0.73	0.76	0.61

Jadual 4.8(a) di atas menunjukkan keputusan peramalan taburan hujan bagi kelompok berlainan. Bagi eksperimen yang melibatkan penggunaan atribut dari kelompok data kaji cuaca yang berlainan sebagai data input kepada pengujian, didapati keputusan peramalan taburan hujan yang dihasilkan memberikan prestasi peramalan yang semakin baik atau meningkat selari dengan pertambahan bilangan atribut data kaji cuaca yang digunakan. Ini ditunjukkan oleh kesemua eskperimen yang melibatkan penggunaan data set (kecuali 100 set data menunjukkan prestasi yang tidak sekata). Jadual 4.8(a) juga menunjukkan prestasi peramalan taburan hujan yang dihasilkan di dalam eksperimen yang melibatkan penggunaan 400 set data memberikan prestasi yang terbaik berbanding dengan eksperimen yang lain.

Jadual 4.8(b): Keputusan peramalan taburan hujan bagi kelompok sama

KELOMPOK	100 Set Data		200 Set Data		300 Set Data		400 Set Data		500 Set Data	
	RMS	CC	RMS	CC	RMS	CC	RMS	CC	RMS	CC
2 Kelompok	0.92	0.04	0.74	0.43	0.81	0.29	0.97	0.34	0.84	0.34
3 Kelompok	0.92	0.16	0.88	0.13	0.84	0.18	0.55	0.67	0.95	0.13
4 Kelompok	0.94	0.10	0.89	0.11	0.88	0.11	1.11	0.05	0.97	0.04
5 Kelompok	0.95	0.05	0.87	0.15	0.89	0.09	1.11	0.24	0.97	0.01
6 Kelompok	0.95	0.03	0.89	0.08	0.89	0.07	1.11	0.06	0.97	0.01

Manakala Jadual 4.8(b) pula menunjukkan hasil peramalan taburan hujan yang dilakukan ke atas kelompok atribut data kaji cuaca yang sama. Berdasarkan kepada keputusan peramalan taburan hujan yang ditunjukkan di dalam jadual di atas, didapati semakin banyak bilangan atribut data kaji cuaca yang digunakan sebagai data input kepada proses pengujian, semakin menurun prestasi peramalan taburan hujan yang

dihasilkan. Prestasi ini ditunjukkan oleh nilai RMS yang semakin tinggi dan juga nilai pekali korelasi yang semakin menghampiri 0 di dalam kesemua eksperimen yang melibatkan 100 set data, 200 set data, 300 set data, 400 set data dan 500 set data kaji cuaca.

Oleh yang demikian, dapatlah dirumuskan bahawa dengan menggunakan algoritma *Average Link* untuk mengelompokkan data kaji cuaca, didapati keputusan peramalan taburan hujan yang dihasilkan oleh kelompok yang berlainan memberikan prestasi yang lebih baik berbanding dengan keputusan peramalan taburan hujan yang dihasilkan oleh kelompok yang sama.

4.3.2.3 Algoritma *Complete Link*

Jadual 4.9 menunjukkan hasil pengelompokan data kaji cuaca yang dilakukan dengan menggunakan algoritma *Complete Link*. Berdasarkan hasil pengelompokan ini, beberapa eksperimen telah dilaksanakan untuk melihat prestasi peramalan taburan hujan yang dihasilkan. Eksperimen ini dijalankan dengan melibatkan penggunaan atribut-atribut data kaji cuaca dari kelompok yang sama dan juga kelompok yang berlainan.

Jadual 4.9: Hasil pengelompokan data kaji cuaca menggunakan algoritma *Complete Link*

Bil. Kelompok	100 Set Data	200 Set Data	300 Set Data	400 Set Data	500 Set Data
2	(a, d, f, g) (b, c, e)	(a, d, f, g) (b, c, e)	(a, c, d) (b, e, f, g)	(a, c, d) (b, e, f, g)	(a, b, c, d) (e, f, g)
3	(a, d, f, g) (b) (c, e)	(a, d, f, g) (b) (c, e)	(a, c, d) (b) (e, f, g)	(a, c, d) (b) (e, f, g)	(a, c, d) (b) (e, f, g)
4	(a, f, g) (b) (c, e) (d)	(a, f, g) (b) (c, e) (d)	(a, c) (b) (d) (e, f, g)	(a, c, d) (b) (e) (f, g)	(a, c) (b) (d) (e, f, g)
5	(a, f, g) (b) (c) (d) (e)	(a, f, g) (b) (c) (d) (e)	(a, c) (b) (d) (e, f) (g)	(a, c) (b) (d) (e) (f, g)	(a, c) (b) (d) (e, f) (g)

6	(a, f) (b) (c) (d) (e) (g)	(a, f) (b) (c) (d) (e) (g)	(a) (b) (c) (d) (e, f) (g)	(a) (b) (c) (d) (e) (f, g)	(a) (b) (c) (d) (e, f) (g)

a – *windvane*, b – *humidity*, c – *energy*, d – *temp*, e – *tension*, f – *radiation*, g - *windspeed*

Jadual 4.10(a) menunjukkan keputusan peramalan taburan hujan yang melibatkan penggunaan atribut data kaji cuaca dari kelompok berlainan. Manakala 4.10(b) pula menunjukkan keputusan peramalan taburan hujan yang menggunakan atribut data kaji cuaca dari kelompok sama sebagai data input kepada eksperimen yang telah dijalankan.

Jadual 4.10(a): Keputusan peramalan taburan hujan bagi kelompok berlainan

KELOMPOK	**100 Set Data**		**200 Set Data**		**300 Set Data**		**400 Set Data**		**500 Set Data**	
	RMS	CC	RMS	CC	RMS	CC	RMS	CC	RMS	CC
2 Kelompok	0.94	0.09	0.87	0.22	0.88	0.15	1.06	0.10	0.96	0.11
3 Kelompok	0.95	0.08	0.83	0.32	0.84	0.28	0.95	0.34	0.94	0.19
4 Kelompok	0.71	0.27	0.73	0.52	0.76	0.48	0.90	0.55	0.91	0.32
5 Kelompok	0.74	0.16	0.71	0.57	0.60	0.71	0.89	0.59	0.76	0.61
6 Kelompok	0.72	0.25	0.71	0.60	0.56	0.75	0.76	0.73	0.76	0.61

Bagi eksperimen yang melibatkan penggunaan atribut dari kelompok yang sama, didapati keputusan peramalan taburan hujan yang dihasilkan menunjukkan semakin banyak bilangan atribut data kaji cuaca yang digunakan sebagai data input kepada proses peramalan, semakin meningkat prestasi peramalan taburan hujan yang dihasilkan. Ini ditunjukkan oleh kesemua eksperimen yang melibatkan 100 set data, 200 set data, 300 set data, 400 set data dan 500 set data, di mana nilai RMS yang dihasilkan semakin menurun manakala nilai pekali korelasinya semakin meningkat. Selain daripada itu, didapati eksperimen yang menggunakan 200 set data, 300 set data dan 400 set data menunjukkan prestasi peramalan yang lebih baik berbanding yang lain.

Jadual 4.10(b): Keputusan peramalan taburan hujan bagi kelompok sama

KELOMPOK	**100 Set Data**		**200 Set Data**		**300 Set Data**		**400 Set Data**		**500 Set Data**	
	RMS	CC	RMS	CC	RMS	CC	RMS	CC	RMS	CC
2 Kelompok	0.95	0.08	0.95	0.09	0.83	0.35	0.91	0.49	0.87	0.41
3 Kelompok	0.94	0.11	0.80	0.33	0.84	0.30	0.90	0.26	0.92	0.19
4 Kelompok	0.95	0.06	0.87	0.22	0.87	0.16	0.95	0.18	0.93	0.14
5 Kelompok	0.95	0.05	0.87	0.17	0.88	0.12	0.95	0.10	0.98	0.04
6 Kelompok	0.95	0.04	0.89	0.10	0.88	0.16	0.95	0.09	0.98	0.03

Berdasarkan keputusan peramalan yang dipaparkan pada Jadual 4.10(b) di atas, didapati prestasi peramalan yang dihasilkan semakin menurun selari dengan peningkatan bilangan atribut data kaji cuaca yang digunakan di dalam proses peramalan taburan hujan tersebut. Ini ditunjukkan oleh kesemua eksperimen yang dijalankan, yang melibatkan penggunaan 100 set data, 200 set data, 300 set data, 400 set data dan 500 set data.

Oleh yang demikian, maka bolehlah dirumuskan bahawa dengan mengelompokkan data kaji cuaca menggunakan algoritma *Complete Link*, didapati prestasi peramalan taburan hujan yang dihasilkan dengan menggunakan atribut data kaji cuaca dari kelompok berlainan adalah lebih baik jika dibandingkan dengan prestasi peramalan taburan hujan yang dihasilkan dengan menggunakan atribut data kaji cuaca dari kelompok sama. Ini dibuktikan oleh nilai RMS dan pekali korelasinya yang ditunjukkan pada Jadual 4.10(a) dan Jadual 4.10(b).

BAB V

KESIMPULAN

Terdapat pelbagai teknik yang boleh digunakan untuk melakukan pengelompokan data, di antaranya ialah teknik perlombongan data, kaedah statistik dan sebagainya. Oleh yang demikian, kajian ini dijalankan bertujuan untuk mengkaji salah satu kaedah pengelompokan yang terdapat dalam teknik perlombongan data iaitu algoritma K-Means. Sehubungan dengan itu, kajian ini telah dilaksanakan dengan mengaplikasikan algoritma K-Means di dalam mengelompokkan data kaji cuaca bagai tujuan peramalan taburan hujan. Hasil daripada eksperimen peramalan taburan hujan yang telah dijalankan ini menunjukkan prestasi peramalan taburan hujan adalah berkadar terus dengan bilangan data kaji cuaca yang digunakan sebagai data input kepada proses peramalan tersebut. Ini bermaksud semakin tinggi jumlah data kaji cuaca yang digunakan, semakin tinggi prestasi peramalan taburan hujan yang dihasilkan (ditunjukkan oleh hasil peramalan bagi semua kelompok 2, 3, 4, 5 dan 6). Selain daripada itu, keputusan eksperimen juga menunjukkan semakin besar bilangan kelompok data kaji cuaca yang digunakan, semakin tinggi prestasi peramalan yang dihasilkan (bagi semua set data 100, 200, 300, 400 dan 500 yang digunakan).

Bagi eksperimen yang telah dijalankan ke atas pengelompokkan hierarchical pula, maka secara keseluruhannya bolehlah dirumuskan bahawa prestasi peramalan taburan hujan yang dihasilkan oleh eksperimen yang melibatkan penggunaan atribut-atribut data kaji cuaca dari kelompok yang berlainan adalah lebih baik jika dibandingkan dengan prestasi peramalan taburan hujan yang menggunakan atribut data kaji cuaca dari kelompok yang sama. Ini ditunjukkan oleh nilai RMS eksperimennya yang semakin menghampiri 0 dan juga nilai pekali korelasinya yang semakin menghampiri 1 di dalam

kesemua eksperimen yang melibatkan ketiga-tiga teknik pengelompokan iaitu algoritma *Single Link, Average Link* dan juga *Complete Link*.

Walau bagaimanapun, didapati algoritma *Complete Link* memberikan prestasi peramalan taburan hujan yang lebih baik berbanding dua algoritma yang lain, diikuti oleh algoritma *Single Link* dan akhir sekali *Average Link*. Sehubungan dengan itu, kajian ini boleh diteruskan dengan membuat perbandingan di antara teknik pengelompokan Hierarki *Agglomorative* dengan teknik-teknik pengelompokan yang lain, contohnya teknik pengelompokan Hierarki *Divisive* (seperti *Divisive Analysis* dan *Monothetic Analysis*) dan teknik pengelompokan *Partitional* (seperti *Square Error, Graph Theoretic, Mixture Resolving* dan *Mode Seeking*).

5.1 Kelebihan dan Kelemahan Kajian

Kajian ini mempunyai beberapa kelebihan dan kelemahan. Di antara kelebihannya ialah ia dapat memberikan satu cadangan atau alternatif baru kepada pihak JPKM untuk melakukan penganalisaan data kajicuaca bagi tujuan peramalan. Manakala kelemahannya pula ialah kajian ini tidak dapat melakukan peramalan taburan hujan untuk masa hadapan.

5.2 Cadangan Kajian Lanjutan

Kajian terhadap teknik pengelompokan data kajicuaca ini merupakan satu bidang yang menarik untuk diselidiki. Ini kerana operasi pengelompokan telah banyak digunakan di dalam pelbagai bidang terutamanya bidang pengelompokan imej (Shahliza, 1999), aksara (Azah Kamilah, 1999) data tak berstruktur jenis teks (Siti Sakira, 1998) dan sebagainya. Dan terdapat banyak teknik pengelompokan terutamanya teknik perlombongan data yang boleh digunakan untuk melakukan pengelompokan.

Dalam kajian ini, perbandingan di antara kaedah statistik dan teknik peraturan kesatuan, iaitu salah satu teknik perlombongan data bagi melakukan operasi pengelompokan telah dilaksanakan. Oleh yang demikian, diharapkan ianya dapat dikembangkan lagi pada masa akan datang. Umpamanya;

i) Menggunakan data kajicuaca dalam bentuk harian atau mingguan.

ii) Menggunakan pakej yang lain bagi tujuan perbandingan peramalan, contohnya pakej NeuralWorks dan lain-lain.

iii) Membuat peramalan taburan hujan pada masa hadapan.

iv) Menggunakan teknik pengelompokan yang lain seperti teknik jiran terdekat.

v) Membandingkan hasil peramalan rangkaian neural dengan kaedah statistik yang lain, contohnya regresi dan lain-lain.

5.3 Sumbangan

Di antara sumbangan kajian ini ialah;

i) Membantu pihak JPKM untuk meramalkan taburan hujan di masa akan datang dengan menggunakan kaedah pengelompokan data kajicuaca yang difikirkan sesuai.

ii) Memberikan sedikit sumbangan terhadap bidang kajian teknik peraturan kesatuan dan kaedah statistik di dalam melakukan pengelompokan data siri masa, khususnya data kajicuaca.

iii) Membuat perbandingan di antara teknik pengelompokan dan juga kelompok yang digunakan untuk meramal taburan hujan.

iv) Memberi pendedahan dan juga pengetahuan kepada orang awam tentang kepentingan dan kegunaan maklumat yang berkaitan dengan kajicuaca dan peramalan taburan hujan.

LAMPIRAN A: (KEPUTUSAN EKSPERIMEN PERTAMA HINGGA EKSPERIMEN KELIMA)

Jadual 4.4(a): Keputusan Peramalan Taburan Hujan Bagi 100 Set Data

KELOMPOK	RMS	CC
2	0.97	0.05
3	0.95	0.18
4	0.95	0.18
5	0.95	0.21
6	0.74	0.46

Jadual 4.4(b): Keputusan Peramalan Taburan Hujan Bagi 200 Set Data

KELOMPOK	RMS	CC
2	0.95	0.15
3	0.87	0.28
4	0.82	0.30
5	0.81	0.33
6	0.74	0.48

Jadual 4.4(c): Keputusan Peramalan Taburan Hujan Bagi 300 Set Data

KELOMPOK	RMS	CC
2	0.94	0.15
3	0.79	0.30
4	0.75	0.36
5	0.70	0.39
6	0.69	0.57

Jadual 4.4(d): Keputusan Peramalan Taburan Hujan Bagi 400 Set Data

KELOMPOK	RMS	CC
2	0.90	0.18
3	0.78	0.34
4	0.73	0.38
5	0.68	0.42
6	0.65	0.65

Jadual 4.4(e): Keputusan Peramalan Taburan Hujan Bagi 500 Set Data

KELOMPOK	RMS	CC
2	0.87	0.20
3	0.76	0.36
4	0.70	0.42
5	0.65	0.50
6	0.60	0.75

LAMPIRAN B

PEMBENTANGAN KERTAS KERJA DI SIMPOSIUM KEBANGSAAN SAINS MATEMATIK 31 MEI – 2 JUN 2005 ALOR STAR, KEDAH

PENGELOMPOKAN DATA KAJI CUACA MENGGUNAKAN K-MEANS BAGI PERAMALAN TABURAN HUJAN.

Mahadi Bahari[1] , Rozilawati Dollah @ Md. Zain[2],
Aryati Bakri[3], Mohamad Fahmi Mohamad Adini[4]

Fakulti Sains Komputer dan Sistem Maklumat
Universiti Teknologi Malaysia
81310 Skudai, Johor.
Tel : 07-5576160 ext. [1]34207, [2]32425, [3]32428
Fax : 07-5565044
E-Mel : {[1]mahadi, [2]zeela, [3]aryati}@fsksm.utm.my,
[3]fahmi_adini@hotmail.com}

Abstrak

Pengelompokan merupakan salah satu teknik utama di dalam perlombongan data di mana set entiti dibahagikan kepada beberapa subkelas. Tujuan utama proses pengelompokan adalah untuk mengenalpasti corak sesebuah kumpulan, yang membolehkan persamaan serta perbezaan yang wujud antara kumpulan dikenalpasti. Terdapat pelbagai kaedah di dalam pengelompokan di mana setiap satunya berfungsi mengikut cara tersendiri dan mengeluarkan keputusan yang berlainan. Kajian ini menekankan kepada proses pengelompokan data kaji cuaca dengan menggunakan algoritma K-Means. Di dalam kajian ini, beberapa algoritma K-Means yang dihasilkan oleh penulis yang berlainan, dibincangkan secara umum. Penulis juga telah memfokuskan kajian kepada pengelompokan parameter data kaji cuaca kepada beberapa kelompok yang berlainan. Kelompok-kelompok ini dihasilkan melalui pembangunan algoritma K-Means dengan menggunakan program Borldan C. Hanya beberapa wakil parameter sahaja (diambil daripada setiap kelompok yang telah dihasilkan), akan digunakan bagi melakukan proses peramalan taburan hujan. Hasil daripada eksperimen peramalan taburan hujan yang telah dijalankan ini menunjukkan prestasi peramalan taburan hujan adalah berkadar terus dengan bilangan data kaji cuaca yang digunakan sebagai data input kepada proses peramalan tersebut.

Kata Kunci: Pengelompokan, K-Means, Peramalan taburan hujan, Kaji cuaca.

1.0 Pengenalan

Keadaan cuaca setempat memberi kesan yang mendalam ke atas hidrologi permukaan bumi [Kim dan Miller(1996)] dan kitaran air di dalam tanah memberi kesan ke atas persekitaran, sumber air dan aktiviti manusia. Salah satu elemen terpenting di dalam cuaca ialah peramalan taburan hujan di mana ia memainkan peranan utama di dalam bidang kaji cuaca bagi pemerhatian cuaca. Selain daripada itu. ia merupakan tugas yang mencabar di dunia sejak lebih daripada setengah dekad yang lalu [Chen dan Takagi(1993); Ultsch dan Guimareas(1996); Liu dan Lee(1999); McCullagh dan rakan-rakan(1999)]. Jabatan Perkhidmatan Kaji Cuaca Malaysia (JPKM) merupakan sebuah agensi yang menjadi pusat pemerhatian perubahan cuaca bagi Negara Malaysia. JPKM juga bertanggungjawab di dalam proses peramalan taburan hujan [JPKM(2004)] bagi mengelakkan sebarang bencana alam seperti banjir atau kemarau yang merupakan dua fenomena alam yang penting dan memberi kesan ke atas kehidupan manusia [McCullagh dan rakan-rakan(1999)]. Fenomena ini juga memberi kesan ke atas ekonomi setempat dan boleh mendatangkan kemudaratan. Peramalan cuaca yang tepat adalah penting bagi membolehkan pemberitahuan amaran awal bencana serta membantu proses pengurusan sumber air [Kim dan Miller (1996)].

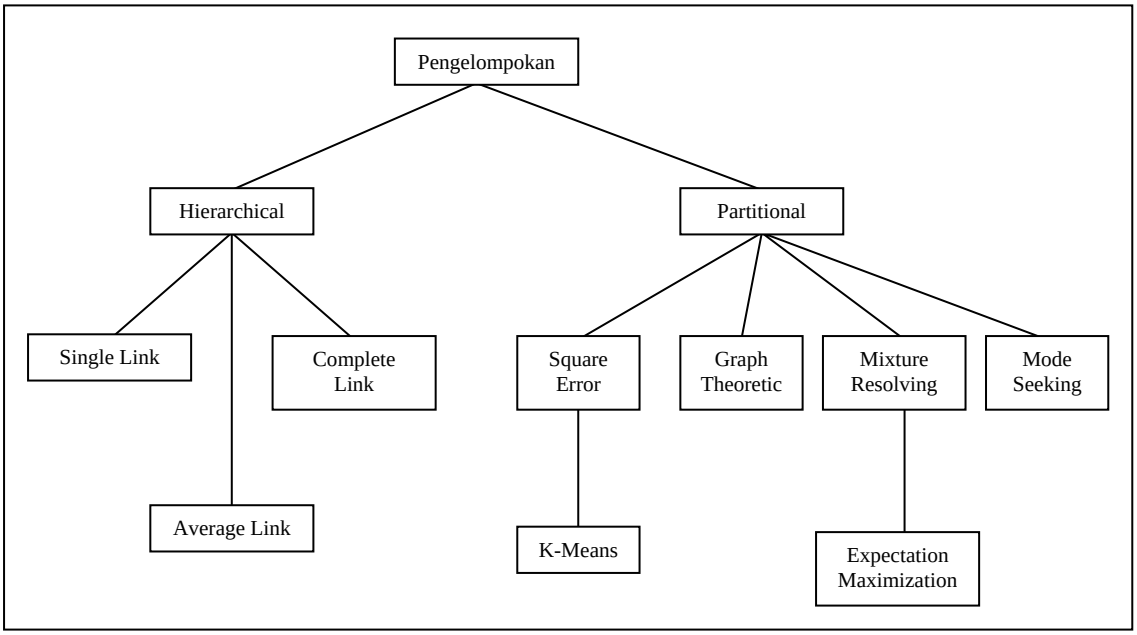
Hujan merupakan salah satu elemen yang rumit serta kompleks untuk diramal kerana ianya mempunyai kepelbagaian pembolehubah atau parameter yang wujud serta mempunyai kaitan yang kompleks antara satu sama lain [Chen dan Takagi(1993); Ultsch dan Guimareas(1996)]. Menurut JPKM, terdapat pelbagai parameter yang mempunyai kaitan dengan corak hujan iaitu *windvane*, *humidity*, *energy*, *temperature*, *tension*, *radiation* dan *windspeed*. Parameter-parameter ini digunakan sebagai input bagi membuat peramalan taburan hujan bagi jangka pendek atau panjang. Kepelbagaian parameter inilah menyebabkan peramalan taburan hujan menjadi satu tugas yang rumit.

Objektif utama kajian ini adalah untuk mengekstrak maklumat yang berguna daripada sejumlah data kaji cuaca yang ada melalui penggunaan kaedah pengelompokan. Di dalam kajian ini, kaedah pengelompokan parameter-parameter data kaji cuaca digunakan untuk memudahkan proses peramalan taburan hujan dibuat. Terdapat pelbagai kajian yang telah dilakukan ke atas peramalan taburan hujan seperti yang dilakukan oleh Diyankov (1992); Chen dan Takagi (1993); Oichiai (1995);

McCullagh (1995;1999); Oishi (1998); Liu dan Lee (1999); Jorge (2000) dan Hui Qi (2001)] yang mengadaptasikan rangkaian neural dalam mengklasifikasikan corak hujan. Walaupun begitu, kajian ini hanya menekankan kepada teknik pengelompokan data menggunakan algoritma K-Means bagi menghasilkan kelompok-kelompok data kaji cuaca untuk digunakan dalam peramalan taburan hujan.

2.0 Teknik Pengelompokan

Pengelompokan merupakan salah satu kaedah yang popular di dalam perlombongan data di mana beberapa set entiti dibahagikan kepada beberapa kumpulan atau subkelas yang bermakna, yang dipanggil kelompok. Setiap elemen yang terdapat di dalam kelompok mempunyai persamaan antara satu sama lain (atau boleh dikategorikan sebagai satu kumpulan) dan terdapat perbezaan antara satu kelompok dengan kelompok yang lain. Tujuan utama proses pengelompokan adalah untuk mengenalpasti corak sesebuah kumpulan, di mana membolehkan kita melihat persamaan serta perbezaan yang wujud antara kumpulan. Ini membolehkan andaian serta peramalan dapat dibuat berdasarkan kumpulan yang telah dikelompokkan ini. Terdapat pelbagai kaedah di dalam pengelompokan data di mana setiap kaedah berfungsi mengikut cara tersendiri dan mengeluarkan keputusan yang berlainan [Zait dan Metsaffa(1997)].



Rajah 1: Teknik Pengelompokan.

Kepelbagaian kaedah di dalam pengelompokan data ditunjukkan pada Rajah 1. Terdapat dua kategori utama di dalam kaedah pengelompokan iaitu *hierarchical* dan *partitional*. Teknik-teknik di dalam kategori *hierarchical* akan mengelompokkan pangkalan data kepada beberapa pembahagian bersarang. Terdapat dua jenis algoritma bagi pengelompokan *hierarchical* iaitu *agglomerative* dan *divisive* [Hirano dan Tsumoto(2004)]. Algoritma *agglomerative* mengumpukkan setiap objek sebagai kelompok. Selepas itu, ia akan mencari pasangan yang mempunyai persamaan ciri dan dikelompokkan sebagai satu kelompok. Proses pencarian ini akan berterusan sehingga kesemua objek telah dimasukkan ke dalam kelompok-kelompok yang berkaitan. Manakala algoritma *divisive* melakukan tugas yang berlawanan dengan *hierarchical*. Ia akan mengumpukkan kesemua objek sebagai satu kelompok. Kelompok-kelompok tadi akan dipecahkan kepada beberapa kelompok yang sama. Algoritma *hierarchical* dibahagikan kepada tiga sub kategori iaitu *single link*, *average link* dan *complete link*. Setiap satu mempunyai kelainan dari segi melakukan pengkategorian persamaan pasangan sesuatu kelompok [Jain dan rakan-rakan (1999)].

Kertas kerja ini hanya menumpukan kepada pengelompokan algoritma *partitional*. Algoritma *partitional* mempunyai kelainan berbanding *hierarchical* di mana ia akan membentuk data kepada k

kelompok. Secara praktiknya, algoritma *partitional* akan diproses beberapa kali dengan berlainan keadaan awalan dan konfigurasi yang terbaik akan dijadikan sebagai output kepada pengelompokan yang telah dijalankan [Jain dan rakan-rakan(1999)]. Seperti yang ditunjukkan di dalam Rajah 1, terdapat empat sub kategori *partitional* iaitu *square error*, *graph theoretic*, *mixture resolving* dan *mode seeking*. Keempat-empatnya akan mengkategorikan data kepada beberapa kelompok yang berlainan. Ia akan mengenalpasti bilangan kelompok yang dapat dijana berdasarkan fungsi kriteria bagi tujuan mengoptimumkan data [Haldiki dan rakan-rakan(2001)]. Fungsi kriteria yang paling kerap digunakan ialah *square error* di mana setiap pengiraan jarak daripada kelompok tengah akan dijumlahkan bagi setiap set data yang terlibat. Ia juga dikenali sebagai algoritma *square error*. Salah satu algoritma yang meminimumkan *square error* ini adalah algoritma K-Means.

3.0 Algoritma K-Means

Algoritma K-Means merupakan satu algoritma yang mudah dan kerap digunakan di dalam teknik pengelompokan kerana ia melibatkan pengiraan yang efisien dan tidak memerlukan banyak parameter. K-Means [MacQueen(1967)] menggunakan k kelompok yang telah ditetapkan (k kelompok pertama sebagai centroid) dan secara berterusan akan melalui proses pengiraan titik tengah (min) sehingga sesuatu fungsi kriteria dicapai (kelompok adalah tetap). Di dalam teknik pengelompokan, pengiraan untuk membezakan di antara kelompok dilakukan menggunakan satu algoritma yang dipanggil fungsi jarak iaitu tahap persamaan atau perbezaan.

Pengukuran persamaan atau jarak merupakan tugas yang penting di dalam proses analisa kelompok di mana hampir semua teknik pengelompokan menggunakan pengiraan matriks jarak (atau perbezaan) [Doherty dan rakan-rakan(2001)]. Algoritma K-Means juga menggunakan kaedah pengiraan ini bagi menjelaskan lagi persamaan bagi setiap corak kelompok. Matriks Jarak Euclidean merupakan salah satu matriks jarak yang sering digunakan di dalam algoritma K-Means.

Matriks Jarak Euclidean

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \qquad (1)$$

di mana $x = (x_1, x_2, ..., x_n)$ dan $y = (y_1, y_2, ..., y_n)$

d (x,y) = jarak di antara x dan y
y_i = nilai pembolehubah $_i$ bagi x
x_i = nilai pembolehubah $_i$ bagi y

Di dalam kajian ini, kita akan melihat algoritma yang dilakukan oleh beberapa penulis seperti yang ditunjukkan di dalam Jadual 1:

Jadual 1: Algoritma K-Means

No.	Penulis	Algoritma
1.	Kim, B. J., Kripalani, R. H., Oh, J. H., dan Moon, S. E. [2002]	(i) Wujudkan k kelas. Pilih secara rambang k corak daripada seluruh set data dan umpukkan setiap set data kepada setiap kelas. Pada fasa ini, min corak data setiap set data mengikut corak. (ii) Umpukkan setiap corak kepada set data kepada kelas di mana min yang terdekat berdasarkan pengukuran jarak $\delta_{(i,j)}$ iaitu; $\delta_{(i,j)} = \sum_{i,j=1}^{m} (X_i - X_j)^2$ (2) (iii) Kira nilai min yang baru bagi setiap kelas. (iv) Ulang langkah (ii) dan semak jika berlaku sebarang perubahan corak pada kelas. Jika ya, ulang langkah (iii) dan (iv).
2.	Al-Harbi, S. H., Rayward-Smith, V. J. [2003]	(i) Pemilihan secara rambang k kelompok, $C_i, 1 \le i \le k$ dan pengiraan centroid bagi setiap kelompok, $\hat{c}_i$.

		(ii)	Kira jarak antara objek dan centroid bagi setiap kelompok.
		(iii)	Umpukkan semula objek pada setiap kelompok.
		(iv)	Ubah centroid bagi setiap kelompok daripada yang telah dibuang dan setiap objek yang telah diumpukkan.
		(v)	Langkah (ii) dan (iv) diulang sehingga kelompok stabil.
3.	Phillips, S. J. [2002]		Anggapkan $u_1,...,u_k$ menjadi min setiap kelas.
		(i)	Umpukkan setiap titik $p \in P$ kepada kelas C_j yang meminimumkan $d(p,u_j)$.
		(ii)	Kira semula min; bagi setiap $j \in \{1...k\}$, set u_j yang menjadi min bagi setiap titik yang diumpukkan C_j di dalam langkah (i).
4.	Wan, S. J., Wong, S. K. M., dan Prusinkiewicz, P. [1988]	(i)	Pilih k kelompok awalan.
		(ii)	K kelompok dibentuk dengan mengumpukkan setiap data kepada kelompok yang terdekat.
		(iii)	Centroid bagi setiap k kelompok menjadi titik tengah yang baru bagi kelompok.
		(iv)	Langkah akan diulang sehingga kelompok baru yang dibentuk sama dengan sebelumnya.
5.	Pena, J. M., Lozana, J. A., dan Larranaga, P. [1999]	(i)	Pilih pembahagian awalan setiap data kepada k kelompok $\{C_1,...,C_k\}$.
		(ii)	Kira centroids $\overline{w}_i = \frac{1}{K_i}\sum_{j=1}^{K_i} w_{ij,} \quad i = 1,...,K$ (3)
		(iii)	Bagi setiap w_i di dalam data dan mengikut susunan objek, Umpukkan objek w_i kepada centroid terdekat, $w_i \in C_s$ dipindahkan daripada C_s kepada C_t jika $\|w_i - \overline{w}_t\| \le \|w_i - \overline{w}_j\|$ bagi setiap $j = 1,...,K, j \ne s$.
		(iv)	Kira semula centroids bagi setiap kelompok C_s dan C_t. Jika data setiap kelompok stabil maka proses diberhentikan. Jika tidak ulang langkah (iii).
6.	Cheung, Y. [2003]	(i)	Umpukkan k kelompok awalan, dan kira nilai asas $\{m_j\}_{j=1}^{k}$. Jika $j = \arg\min_{1\le r\le k} \|x_t - m_r\|^2;$ (4)
		(ii)	Diberi input x_t, kira $I(j \mid x_t)\{_{0 \quad otherwise}^{1 \quad \text{If } j=\arg\min_{1\le r\le k}\|x_t-m_r\|;}$ (5)
		(iii)	Kemaskini nilai *winning seed point* m_w ,melalui $m_w^{new} = m_w^{old} + \eta(x_t - m_w^{old})$, (6)
		(iv)	Di mana η merupakan *small positive learning rate.* Langkah (ii) dan (iii) bagi setiap input.
7.	Bdanyopadhyay, S., dan Maulik, U. [2002]	(i)	Pilih k kelompok awalan $z_1, z_2,..., z_K$ secara rambang daripada n data $\{x_1, x_2,..., x_n\}$.
		(ii)	Umpukkan data $x_i, i = 1,2,...,n$ kepada kelompok C_j, $j \in \{1,2,...,K\}$ jika $\|x_i - z_j\| \le \|x_i - z_p\|,$ (7) $p = 1,2,...,K,$ dan $j \ne p$.
		(iii)	Kira kelompok $z_1^*, z_2^*,..., z_K^*$, seperti berikut:

		$z_i^* = \frac{1}{n_i} \sum_{x_j \in c_i} x_j$, (8) $i = 1,2,...,K$, (iv) Di mana n_i merupakan elemen bagi kelompok C_i . Jika $z_i^* = z_i \forall i = 1,2,...,K$ maka proses diberhentikan. Selain itu ulang langkah (ii).
8.	Smith, K. A., dan Ng, A. [2003]	(i) Nilai awalkan k kelompok sebagai kelompok tengah (guna k kelompok pertama sebagai asas). (ii) Umpukkan setiap data kepada kelompoknya yang terhampir (pengiraan daripada kelompok tengah). Ini dilakukan oleh setiap data x dan pengiraan persamaan (jarak) d melalui input ini kepada berat, w bagi setiap kelompok tengah, j . Kelompok tengah yang terhampir dengan set data x ialah kelompok tengah dengan jarak minimum dengan data x . $d_j = \left\| x - w_j \right\| = \sqrt{\sum_{i=1}^{n} (x_i - w_{ij})}$ (9) (iii) Kira semula titik tengah bagi setiap kelompok sebagai centroid bagi setiap set data dalam setiap kelompok. Centroid ρc dikira seperti berikut: $\rho c = < w_1^c, w_2^c, ..., w_n^c >$ (10) Di mana $w_1^c = \frac{\sum_{j \in c} u_i^j}{N^c}$ (11) Di mana : N^c merupakan bilangan data di dalam kelompok. (iv) Jika kelompok tengah baru adalah berlainan dengan sebelumnya, ulang langkah (ii). Jika tidak, proses diberhentikan.

Berdasarkan Jadual 1 di atas, dapat disimpulkan bahawa tujuan utama K-Means ialah mengenalpasti k kelompok sebagai centroid dan mengumpukkan data kepada centroid yang terhampir (sama). Pada tahap ini, pengiraan semula k kelompok baru berdasarkan hasil sebelumnya. Selepas mendapat k kelompok yang baru, pengiraan kepada centroid yang terdekat perlu dilakukan ke atas semua set data. Proses semakan akan dilakukan bagi memastikan setiap set data menepati persamaan dengan centroid. Proses ini akan berulang sehingga tidak berlaku perubahan ke atas lokasi centroid atau dengan kata lain, tidak ada perbezaan lagi antara set data dengan centroid. Di dalam kajian ini, penggunaan algoritma K-Means diambil daripada penulis Smith, K. A., dan Ng, A. [2003].

4.0 Metodologi

Kajian ini dilaksanakan mengikut beberapa aktiviti. Antara aktiviti-aktiviti yang terlibat di dalam perlaksanaan kajian ini adalah seperti berikut :-

i) **Mengumpul dan menganalisa data kaji** cuaca : perlaksanaan kajian ini melibatkan penggunaan 100, 200, 300, 400 dan 500 set data kaji cuaca yang diperolehi daripada Jabatan Perkhidmatan Kaji cuaca Malaysia bermula dari 1 Ogos 2000 sehingga 21 Ogos 2000. Ia mengandungi lapan atribut data kaji cuaca yang digunakan, iaitu atribut *windvane, humidity, energy, temp, tension, radiation, windspeed* dan *rainfall*.

ii) **Pengiraan jarak Euclidean** : jarak Euclidean digunakan untuk mengira persamaan di antara atribut-atribut data kaji cuaca bagi mengenalpasti atribut-atribut yang mempunyai persamaan yang kuat dan lemah untuk tujuan pengelompokan. Formula untuk pengiraan jarak Euclidean ialah;

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \qquad (12)$$

di mana $x = (x_1, x_2, ..., x_n)$ dan $y = (y_1, y_2, ..., y_n)$

iii) **Pembangunan aturcara** : selepas formula pengiraan persamaan antara kelompok dikenalpasti, algoritma K-Means dibangunkan dengan menggunakan aturcara program Borland C. Algoritma K-Means digunakan untuk mengelompokkan data kaji cuaca menggunakan jarak Euclidean untuk mengasingkan set-set data kaji cuaca kepada beberapa kelompok yang berkaitan.

iv) **Pengelompokan set data kaji cuaca** : berdasarkan kepada hasil pengiraan jarak melalui aturcara yang telah dibangunkan, proses pengelompokan data kaji cuaca dilakukan di mana, atribut-atribut yang mempunyai jarak terkecil dikira sebagai atribut yang mempunyai persamaan yang tinggi dan akan dikelompokkan ke dalam satu kelompok yang sama. Manakala bagi atribut-atribut yang mempunyai jarak yang besar dikira sebagai atribut yang mempunyai persamaan yang rendah dan dikelompokkan ke dalam kelompok yang berlainan. Proses pengelompokan ini dilakukan beberapa kali di mana ia melibatkan penghasilan 2 kelompok, 3 kelompok, 4 kelompok, 5 kelompok dan 6 kelompok.

v) **Pengujian kelompok** : setelah pengelompokan dijalankan ke atas set data kaji cuaca, pengujian kelompok pula dilaksanakan dengan melakukan peramalan taburan hujan, menggunakan Pakej NeuNet Pro 2.3. Pengujian ini dilakukan dengan menggunakan atribut-atribut yang terdapat di dalam kelompok yang berlainan sebagai data input kepada proses peramalan taburan hujan. Sehubungan dengan itu, beberapa eksperimen yang melibatkan atribut dari kelompok data kaji cuaca yang yang berlainan telah dilaksanakan. Pengujian kelompok dilakukan untuk membuat penganalisaan dan perbandingan prestasi peramalan taburan hujan di antara set-set eksperimen yang dijalankan.

vi) **Penganalisaan dan perbandingan hasil** : proses pengujian kelompok akan menghasilkan satu keputusan peramalan di antara kelompok-kelompok data kaji cuaca yang berlainan. Penganalisaan akan dilakukan ke atas hasil pengujian untuk menentukan ketepatan teknik pengelompokan menggunakan algoritma K-Means. Penganalisaan ini dilakukan dengan melihat nilai perbezaan di antara nilai sebenar taburan hujan serta nilai ramalan taburan hujan yang dihasilkan. Selain daripada itu, elemen-elemen lain yang turut digunakan sebagai perbandingan ialah nilai ralat min punca kuasa dua (RMS) dan pekali korelasi bagi menentukan keberkesanan algoritma K-Means.

5.0 Eksperimen

Kajian ini melibatkan penggunaan salah satu teknik pengelompokan data, iaitu algoritma K-Means untuk menghasilkan set-set kelompok data kaji cuaca. Penentuan set kelompok data kaji cuaca ini dilakukan berdasarkan kepada pengiraan jarak Euclidean yang telah diaplikasikan dalam aturcara yang telah dibangunkan. Sehubungan dengan itu, proses pembahagian atau penentuan set kelompok ini dilakukan sebanyak lima(5) kali, di mana ia melibatkan 100, 200, 300, 400 dan 500 set data kaji cuaca. Jadual 1 di bawah menunjukkan sampel data kaji cuaca yang digunakan di dalam eksperimen ini.

Jadual 1: Sampel Data Kaji Cuaca

ID	Tarikh	Masa	Windvane	Humidity	Energy	Temp	Tension	Radiation	Windspeed	Rainfall
1	8/1/00	0:00	197	0	-0.58	-10.8	-4.8	0	0	0
2	8/1/00	1:00	201.2	0	-0.63	-11	-4.8	-48.828	0	0

3	8/1/00	2:00	206.5	0	-0.63	-11.1	-5.8	-48.828	0	0
4	8/1/00	3:00	235.5	0	-0.24	-11.4	-5.8	-48.828	0	0
5	8/1/00	4:00	293.7	0	-0.39	-10.9	-4.8	-48.828	0	0
6	8/1/00	5:00	143.2	0	-0.14	-7.8	-4.8	-48.828	0	0
7	8/1/00	6:00	201.2	0	0.04	-6.7	-3.9	-48.828	0	0
8	8/1/00	7:00	343.8	0	0.53	-6.8	0.9	-48.828	0	2.5
9	8/1/00	8:00	261.2	0	42.23	-7.6	0.9	0	0	0.5
10	8/1/00	9:00	304.7	0	58.88	-9.1	-2.9	97.656	0	0
…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…
500	8/21/00	19:00	290.0	1.5	0.04	18.2	0	0	0	0

Bagi setiap kategori (100, 200, 300, 400 dan 500) set data kaji cuaca tersebut, proses pengelompokan ini dilakukan berulangkali bagi menghasilkan kelompok 2, kelompok 3, kelompok 4, kelompok 5 dan kelompok 6. Hasil penentuan kelompok-kelompok ini boleh dirujuk pada **Jadual 2** berikut.

Jadual 2: Penentuan Kelompok Data Kaji Cuaca

Bil. Kelompok	**100 set**	**200 set**	**300 set**	**400 set**	**500 set**
2	(a) (b, c, d, e, f, g)	(a) (b, c, d, e, f, g)	(a) (b, c, d, e, f, g)	(a) (b, c, d, e, f, g)	(a) (b, c, d, e, f, g)
3	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)
4	(a) (e) (c, f) (b, d, g)	(a) (e) (c, f) (b, d, g)	(a) (e) (c, f) (b, d, g)	(a) (e) (c, f) (b, d, g)	(a) (b, e, g) (c, f) (d)
5	(a) (b, g) (c, f) (d) (e)	(a) (b, g) (c, f) (d) (e)	(a) (b, g) (c, f) (d) (e)	(a) (b) (c, f) (d, g) (e)	(a) (b) (c, f) (d, g) (e)
6	(a) (b, g) (c) (d) (e) (f)	(a) (b, g) (c) (d) (e) (f)	(a) (b, g) (c) (d) (e) (f)	(a) (b) (c) (d, g) (e) (f)	(a) (b) (c) (d, g) (e) (f)

Di mana;
a – *windvane*, b – *humidity*, c – *energy*, d – *temp*, e – *tension*, f – *radiation*, g - *windspeed*

Setelah hasil penentuan kelompok-kelompok data kaji cuaca telah dilakukan, pengujian terhadap kelompok-kelompok tersebut pula dilaksanakan. Ini bertujuan untuk melihat keberkesanan algoritma K-Means di dalam mengelompokkan data kaji cuaca. Pengujian terhadap hasil pengelompokan ini dilakukan dengan menggunakan atribut-atribut data kaji cuaca yang berada di dalam kelompok yang berlainan sebagai data input untuk melakukan proses peramalan taburan hujan. Proses peramalan taburan hujan ini dilakukan dengan menggunakan pakej perisian NeuNetPro.

Oleh yang demikian, beberapa eksperimen peramalan taburan hujan telah dilaksanakan, di mana ia bertujuan untuk melihat ketepatan hasil peramalan taburan hujan tersebut. Perlaksanaan eksperimen ini dilakukan dengan menggunakan jumlah set data kaji cuaca yang sama bagi bilangan kelompok yang berbeza untuk meramal taburan hujan. Untuk tujuan tersebut, terdapat lima(5) set eksperimen telah dikenalpasti dan dilaksanakan, di antaranya;

i) Eksperimen Pertama – melibatkan 100 set data kaji cuaca bagi kelompok 2, 3, 4, 5 dan 6.
ii) Eksperimen Kedua – melibatkan 200 set data kaji cuaca bagi kelompok 2, 3, 4, 5 dan 6.
iii) Eksperimen Ketiga - melibatkan 300 set data kaji cuaca bagi kelompok 2, 3, 4, 5 dan 6.
iv) Eksperimen Keempat - melibatkan 400 set data kaji cuaca bagi kelompok 2, 3, 4, 5 dan 6.
v) Eksperimen Kelima - melibatkan 500 set data kaji cuaca bagi kelompok 2, 3, 4, 5 dan 6.

Bagi eksperimen yang melibatkan kelompok 2, penulis telah melaksanakan enam eksperimen yang berasingan di mana ia melibatkan penggunaan atribut (a,b), atribut (a,c), atribut (a,d), atribut (a,e), atribut (a,f) dan atribut (a,g) sebagai data input untuk melakukan proses peramalan taburan hujan. Berdasarkan nilai RMS dan pekali korelasi yang dihasilkan di dalam keenam-enam eskperimen

tersebut, purata bagi nilai RMS dan pekali korelasi dikira dan ambil sebagai nilai RMS dan pekali korelasi bagi kelompok 2.

Manakala bagi eksperimen yang melibatkan kelompok 3, penulis telah melaksanakan lapan eksperimen berasingan yang melibatkan atribut (a,b,c), atribut (a,d,c), atribut (a,e,c), atribut (a,g,c), atribut (a,b,f), atribut (a,d,f), atribut (a,e,f) dan atribut (a,g,f) sebagai data input untuk melakukan peramalan. Kemudian, purata nilai RMS dan pekali korelasi dikira bagi kesemua eksperimen tersebut dan diambil sebagai nilai RMS dan pekali korelasi bagi kelompok 3.

Seterusnya, enam eksperimen bagi kelompok 4 pula dilaksanakan di mana ia melibatkan atribut (a,e,c,b), atribut (a,e,c,d), atribut (a,e,c,g), atribut (a, e,f,b), atribut (a,e,f,d) dan atribut (a,e,f,g). Purata nilai RMS dan pekali korelasinya dikira dan diambil sebagai nilai RMS dan pekali korelasi bagi kelompok 4. Empat eksperimen berikutnya pula dijalankan di mana ia melibatkan kelompok 5. Oleh yang demikian, atribut (a,b,c,d,e), atribut (a,g,c,d,e), atribut (a,b,f,d,e) dan atribut (a,g,f,d,e) telah digunakan sebagai data input kepada proses peramalan taburan hujan. Dan purata nilai RMS dan pekali korelasinya telah diambil sebagai nilai RMS dan pekali korelasi bagi kelompok 5.

Dan akhir sekali, dua eksperimen bagi kelompok 6 telah dilakukan di mana ia melibatkan atribut (a,b,c,d,e,f) dan atribut (a,g,c,d,e,f) sebagai input kepada proses peramalan taburan hujan. Kemudian, purata nilai RMS dan pekali korelasinya telah diambil sebagai nilai RMS dan pekali korelasi bagi kelompok 6. Jadual 3(a), 3(b), 3(c), 3(d) dan 3(e) pada Lampiran A menunjukkan keputusan peramalan taburan hujan yang dihasilkan di dalam Eksperimen Pertama hingga Eksperimen Kelima di atas.

6.0 Perbincangan

Di dalam kajian ini, eksperimen peramalan taburan hujan yang telah dijalankan adalah bertujuan untuk melihat keupayaan algoritma K-Means di dalam memberikan nilai ramalan yang tepat ke atas data taburan hujan. Berdasarkan keputusan peramalan yang telah dihasilkan tersebut, didapati keputusan peramalan taburan hujan yang dihasilkan menunjukkan semakin besar bilangan kelompok data kaji cuaca yang digunakan sebagai data input kepada proses peramalan, semakin tinggi prestasi peramalan taburan hujan yang dihasilkan, bagi semua set data (jumlah berlainan) yang digunakan. Ini dibuktikan oleh nilai RMS yang semakin berkurangan dan juga nilai pekali korelasi yang semakin tinggi (bagi kelompok 2, 3, 4, 5, dan 6). Selain daripada itu, hasil peramalan juga menunjukkan semakin banyak jumlah data set kaji cuaca yang digunakan untuk melakukan peramalan taburan hujan, semakin tinggi prestasi peramalan yang dihasilkan. Prestasi peramalan ini ditunjukkan oleh keputusan nilai RMSnya yang semakin berkurangan, manakala nilai pekali korelasinya yang semakin meningkat bagi semua kelompok 2, 3, 4, 5 dan 6.

Walau bagaimanapun, terdapat beberapa masalah yang timbul semasa perlaksanan eksperimen ini, di antaranya ialah;

i) Sebahagian daripada data-data kaji cuaca yang digunakan di dalam kajian ini adalah data melampau (data melampau bermaksud julat antara data yang berturutan adalah bersaiz besar). Oleh yang demikian, ini telah menyebabkan hasil peramalan taburan hujan kurang memuaskan.
ii) Masalah kelemahan pakej NeuNetPro yang digunakan untuk pengujian peramalan taburan hujan. Ini kerana pakej tersebut tidak dapat digunakan untuk membuat peramalan masa hadapan. Selain daripada itu, pakej ini juga tidak dapat menjanakan proses peramalan jika bilangan data input yang digunakan, kurang daripada 10 data.

Di samping itu, penulis juga telah membuat andaian bagi menjayakan eksperimen yang telah dijalankan di dalam kajian ini. Di antaranya ialah data-data kaji cuaca yang digunakan di dalam eksperimen ini dianggap bersih dan bebas dari hingar. Selain daripada itu, jarak di antara atribut-atribut data kaji cuaca yang paling kecil dianggap mempunyai ciri-ciri persamaan yang kuat dan sebaliknya.

7.0 Kesimpulan

Terdapat pelbagai teknik yang boleh digunakan untuk melakukan pengelompokan data, di antaranya ialah teknik perlombongan data, kaedah statistik dan sebagainya. Oleh yang demikian, kajian ini dijalankan bertujuan untuk mengkaji salah satu kaedah pengelompokan yang terdapat dalam teknik perlombongan data iaitu algoritma K-Means. Sehubungan dengan itu, kajian ini telah dilaksanakan dengan mengaplikasikan algoritma K-Means di dalam mengelompokkan data kaji cuaca bagai tujuan peramalan taburan hujan. Hasil daripada eksperimen peramalan taburan hujan yang telah dijalankan ini menunjukkan prestasi peramalan taburan hujan adalah berkadar terus dengan bilangan data kaji cuaca yang digunakan sebagai data input kepada proses peramalan tersebut. Ini bermaksud semakin tinggi jumlah data kaji cuaca yang digunakan, semakin tinggi prestasi peramalan taburan hujan yang dihasilkan (ditunjukkan oleh hasil peramalan bagi semua kelompok 2, 3, 4, 5 dan 6). Selain daripada itu, keputusan eksperimen juga menunjukkan semakin besar bilangan kelompok data kaji cuaca yang digunakan, semakin tinggi prestasi peramalan yang dihasilkan (bagi semua set data 100, 200, 300, 400 dan 500 yang digunakan).

Penghargaan

Penulis ingin merakamkan penghargaan kepada Jabatan Perkhidmatan Kaji Cuaca Malaysia (JKPM) Kluang, Johor di atas penggunaan data dan pihak Research Management Centre (RMC) Universiti Teknologi Malaysia di atas sokongan gran Jangka Pendek bagi menjayakan projek penyelidikan ini.

Rujukan

Al-Harbi, S. H., Rayward-Smith, V. J. (2003). The use of a supervised k-means algorithm on real-valued data with applications in health. *IEA/AIE*. 575-581.

Bdanyopadhyay, S., dan Maulik, U. (2002). An evolutionary technique based on k-means algorithm for optimal clustering in R. *Information Science*. 146:221-237.

Chen, T., dan Takagi, M. (1993). Rainfall prediction of geostationary Meteorological satellite images using artificial neural network. *International Geoscience dan Remote Sensing Symposium*. 3:1247-1249.

Cheung, Y. (2003). K*-means : A new generalized k-means clustering algorithm. *Pattern Recognition Letters*. 24(15):2883-2893.

Doherty, K.A.J., Adams, R.G., Davey, N. (2001). Non-Euclidean Norms dan Data Normalization.

Dunham, M.H. (2002). *Data Mining Introductory dan Advanced Topics*. Upper Saddle River, New Jersey.

Fred, A., dan Jain, A. K. (2002). Evidence accumulation clustering based on the k-means algorithm. *Structural, Syntactic, dan Statistical Pattern Recognition, LNCS*. 2396:442-451.

Ganguly, A. R. (2002). A hybrid approach to improving rainfall forecasts. *Computing in Science dan Enfineering*. 4(4):14-21.

Haldiki, M., Batistakis, Y., dan Vazirgiannis, M. (2001). Clustering algorithms dan validity measures. Tutorial paper, *Proceedings of SSDBM Conference*.3-22.

Hartigan, J. A. (1975). *Clustering Algorithms*. Canada. 1-14.

Hirano, S., Sun, X., dan Tsumoto, S. (2004). Comparison of Clustering Methods for Clinical Database. *Information Science*. 159(2):155-165.

Jain, A. K., Murty, M. N., dan Flynn, P. J. (1999). Data Clustering : A review. *ACM Computing Surveys (CSUR)*. 31(3):264-323.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C., Silverman, R., dan Wu, A. Y. (2001). The analysis of a simple k-means clustering algorithm. *Symposium on Computational Geometry*. 100-109.

Kim, B. J., Kripalani, R. H., Oh, J. H., dan Moon, S. E. (2002). Summer monsoon rainfall patterns over South Korea dan associated circulation features. *Theoretical dan Applied Climatology*. 72:65-74.

Kim, J., dan Miller, N. L. (1996). Simulating Winds dan Floods : Regional weather-river prediction dan regional climate research. *IEEE Potentials*. 15(4):17-20.

Kulkarni, A., dan Kripalani, R. H. (1998). Rainfall patterns over India : Classification with Fuzzy C-means method. *Theoretical dan Applied Climatology*. 59:137-146.

Lin, H. (1999). Survey dan implementation of clustering algorithms. Theses. Hsinchu, Taiwan, Republic of China.

Liu, J. N. K., dan Lee, R. S. T. (1999). Rainfall forecasting from multiple point sources using neural networks. *In Proceedings of the 1999 IEEE International Conference on Systems, Man, dan Cybernetics (SMC '99)*. 3:429-434.

Malaysia Meteorological Services (2004).
[Online] Available : http://www.kjc.gov.my/

McCullagh, J., Bluff, K., dan Ebert, E. (1995). A Neural network model for rainfall estimation. *The Second New Zealdan International Two Stream Conference on Artificial Neural Networks dan Expert Systems.* 389-392.

McCullagh, J., Bluff, K., dan Hendtlass, T. (1999). Evolving expert neural networks for meteorological rainfall estimations. Proceedings of the *International Conference on Neural Information Processing dan Intelligent Information Systems IEEE (ICONIP '99).* 2:585-590.

Ochiai, K., Suzuki, H., Shinozawa, K., Fujii, M. dan Sonehara, N. (1995). Snowfall dan rainfall forecasting from weather radar images with artificial neural networks. *Proceedings of IEEE International Conferenc*e. 2:1182-1187.

Pena, J. M., Lozana, J. A., dan Larranaga, P. (1999). An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recognition Letters*. 20(10):1027-1040.

Phillips, S. J. (2002). Accelaration of k-means dan related clustering algorithm. *Revised Papers from the 4th International Workshop on Algorithm Engineering dan Experiments*. 166-177.

Smith, K. A., dan Ng, A. (2003). Web page clustering using a self-organizing map of user navigation patterns. *Decisions Support Systems*. 35:245-256.

Tarsitano, A. (2003). A computational study of several relocation methods for k-means algorithms. *Pattern Recognition Letters*. 36(12):2955-2966.

Ultsch, A., dan Guimareas, G. (1996). Classification dan prediction of hail using self-organizing neural networks. *In Proceedings of the International Conference on Neural Networks ICNN '96.* 1622-1627.

Wan, S. J., Wong, S. K. M., dan Prusinkiewicz, P. (1988). An algorithmfor multidimensional data clustering. *ACM Transactions on Mathematical Software.* 14(4):153-162.

Zait, M., Messatfa, H. (1997). A Comparative Study of Clustering Methods. *Future Generation Computer Systems*, 13:149-159.

LAMPIRAN A (KEPUTUSAN EKSPERIMEN PERTAMA HINGGA EKSPERIMEN KELIMA)

Jadual 3(a) : Keputusan Peramalan Taburan Hujan Bagi 100 Set Data

KELOMPOK	RMS	CC
2	0.97	0.05
3	0.95	0.18
4	0.95	0.18
5	0.95	0.21
6	0.74	0.46

Jadual 3(b) : Keputusan Peramalan Taburan Hujan Bagi 200 Set Data

KELOMPOK	RMS	CC
2	0.95	0.15
3	0.87	0.28
4	0.82	0.30
5	0.81	0.33
6	0.74	0.48

Jadual 3(c) : Keputusan Peramalan Taburan Hujan Bagi 300 Set Data

KELOMPOK	RMS	CC
2	0.94	0.15
3	0.79	0.30
4	0.75	0.36
5	0.70	0.39
6	0.69	0.57

Jadual 3(d) : Keputusan Peramalan Taburan Hujan Bagi 400 Set Data

KELOMPOK	RMS	CC
2	0.90	0.18
3	0.78	0.34
4	0.73	0.38
5	0.68	0.42
6	0.65	0.65

Jadual 3(e) : Keputusan Peramalan Taburan Hujan Bagi 500 Set Data

KELOMPOK	RMS	CC
2	0.87	0.20
3	0.76	0.36
4	0.70	0.42
5	0.65	0.50
6	0.60	0.75

LAMPIRAN C:

PENERBITAN KERTAS KEJA DI

JURNAL TEKNOLOGI MAKLUMAT

FAKULTI SANIS KOMPUTER DAN SISTEM MAKLUMAT, UTM

JILID 17, BIL 2 (DISEMBER 2005)

PENGELOMPOKAN DATA KAJI CUACA MENGGUNAKAN TEKNIK PENGELOMPOKAN HIERARKI AGGLOMERATIVE BAGI PERAMALAN TABURAN HUJAN.

Mahadi Bahari[1] , Rozilawati Dollah @ Md. Zain[2],
Mohd Noor Md Sap[3], Mohamad Fahmi Mohamad Adini[4]

Fakulti Sains Komputer dan Sistem Maklumat
Universiti Teknologi Malaysia
81310 Skudai, Johor.
Tel : 07-5576160 ext. [1]34207, [2]32425, [3]32419
Fax : 07-5565044
E-Mel : {[1]mahadi, [2]zeela, [3]mohdnoor}@fsksm.utm.my,
[4]fahmiadini@gmail.com}

Abstrak

Kertas kerja ini melaporkan penggunaan teknik pengelompokan hierarki *Agglomerative* bagi melakukan peramalan taburan hujan. Tujuan utama kajian ini adalah untuk melihat keberkesanan serta prestasi algoritma yang terdapat di dalam teknik pengelompokan hierarki. Kertas kerja ini bermula dengan penerangan ke atas pengelompokan hierarki yang memfokus kepada algoritma *Single Link*, *Average Link* dan *Complete Link*. Melalui penggunaan algoritma-algoritma tersebut, kelompok dihasilkan berdasarkan pembentukan susunan skema pengelompokan dengan mengurangkan jumlah kelompok bagi setiap proses. Kelompok yang dihasilkan, diperolehi daripada gabungan kelompok-kelompok yang terhampir (sama) kepada satu kelompok. Kelompok-kelompok yang dihasilkan melalui ketiga-tiga algoritma tersebut akan digunakan sebagai input bagi melakukan peramalan taburan hujan. Langkah-langkah yang terlibat di dalam proses pengelompokan ini akan diterangkan dengan lebih jelas di dalam bahagian metodologi kajian. Seterusnya, kertas kerja ini akan menerangkan mengenai eksperimen yang dilakukan ke atas kelompok-kelompok yang dihasilkan dengan menggunakan ketiga-tiga algoritma di atas. Pengukuran prestasi pengelompokan dibuat berdasarkan hasil pengelompokan ialah nilai ralat min punca kuasa dua (RMS) dan nilai pekali kolerasi yang dihasilkan di dalam setiap eksperimen yang telah dijalankan. Hasil kajian menunjukkan bahawa peramalan taburan hujan yang terbaik diperolehi melalui penggunaan algoritma *Complete-Link*.

Kata kunci: Perlombongan data, Pengelompokan, Pengelompokan Hierarki *Agglomerative* dan Peramalan taburan hujan.

1.0 Pengenalan

Peramalan cuaca merupakan suata proses atau kerja yang penting dan rumit dalam sesebuah negara. Keadaan cuaca boleh mempengaruhi keadaan persekitaran seperti keadaan muka bumi, aliran sungai dan bencana alam. Faktor-faktor ini juga memberi kesan kepada keadaan semasa sesebuah negara seperti keadaan ekonomi serta kependudukan masyarakat. Jabatan Perkhidmatan Kaji cuaca Malaysia (JPKM) merupakan agensi terpenting yang memainkan peranan utama dalam mengawal serta menguruskan dan membuat pemerhatian tentang keadaan cuaca di Malaysia.

Elemen yang utama dalam peramalan cuaca ialah peramalan taburan hujan, di mana hujan merupakan satu elemen yang rumit serta kompleks untuk diramal kerana ianya mempunyai kepelbagaian pemboleh ubah yang wujud serta mempunyai kaitan yang kompleks antara satu sama lain (Chen and Takagi, 1993; Ultsch and Guimareas, 1996). Berdasarkan JPKM, pemboleh ubah-pemboleh ubah atau atribut-atribut yang terlibat dengan peramalan taburan hujan ialah *windvane*, *humidity*, *energy*, *temperature*, *tension*, *radiation* dan *windspeed*. Kepelbagaian pemboleh ubah ini menyebabkan tugas peramalan menjadi kompleks.

Sehubungan dengan itu, satu penyelesaian telah dikaji bagi mengatasi masalah peramalan taburan hujan ini. Antara kaedah penyelesaian yang telah dikenalpasti ialah teknik perlombongan data iaitu dengan mengaplikasikan kaedah pengelompokan ke atas data yang ingin dikaji. Matlamat utama kajian yang dijalankan ialah untuk mengurangkan bilangan pembolehubah yang boleh digunakan untuk melakukan peramalan taburan hujan. Kaedah yang digunakan ialah dengan menggunakan teknik pengelompokan hierarchical, di mana perkaitan antara pemboleh ubah akan dikaji bagi membolehkan penggunaan pemboleh ubah-pemboleh ubah yang tertentu sahaja dalam membuat sesuatu peramalan.

Namun begitu sudah terdapat beberapa kajian terperinci yang telah dijalankan ke atas peramalan taburan hujan menggunakan teknik perlombongan data. Kebanyakan kajian yang dilakukan menumpukan kepada penggunaan rangkaian neural sebagai asas peramalan taburan hujan. Dalam kajian yang dilakukan oleh Diyankov et al (1992) berkaitan penggunaan rangkaian neural dalam peramalan cuaca telah menggunakan teknik pengesanan imej di mana ia telah menghasilkan corak imej bagi cuaca sepanjang tahun yang boleh digunakan dalam penentuan peramalan cuaca. Menurut kajian beliau, kajian itu telah mencapai kebarangkalian yang hampir tepat kepada data sebenar (91%). Begitu juga kajian yang dijalankan oleh Chen dan Takagi (1993), di mana kajian mereka tertumpu kepada penggunaan imej satelit dalam

meramal hujan menggunakan rangkaian neural. Penganalisisan yang dijalankan telah menghasilkan ketepatan klasifikasi terhadap pengujian data iaitu 90.45%.

2.0 Teknik Pengelompokan Hierarki Agglomerative

Pengelompokan merupakan salah satu kaedah yang digunakan di dalam perlombongan data. Tujuan utama teknik pengelompokan ialah untuk mengumpulkan elemen-elemen yang mempunyai persamaan ke dalam satu kelompok, di mana setiap kelompok mempunyai perbezaan antara satu sama lain. Pengelompokan membolehkan sesuatu corak atau susunan objek yang berkaitan dapat dikenalpasti. Ini membolehkan sebarang persamaan atau perbezaan di dalam satu set data diketahui. Teknik pengelompokan dibahagikan kepada dua jenis utama iaitu *partitional* dan hierarki. Dalam kajian ini, kami menumpukan kepada pengelompokan hierarki dalam menghasilkan kelompok yang dikehendaki. Pengelompokan hierarki dilakukan dengan menentukan persamaan di antara kelompok untuk digabungkan atau dipecahkan bagi menghasilkan struktur hierarki berdasarkan kepada matriks *proximity*. Hasil daripada pengelompokan hierarki selalunya digambarkan melalui pepohon perduaan atau dendogram (Xu dan Wunsch, 2005). Penggabungan atau pemisahan kelompok dilakukan berdasarkan algoritma di dalam pengelompokan hierarki yang digunakan.

Sejak kebelakangan ini, terdapat pelbagai kajian yang telah dijalankan menggunakan teknik pengelompokan hierarki. Namun begitu, tidak terdapat kajian yang menggunakan teknik pengelompokan hierarki ke atas data kaji cuaca dalam peramalan taburan hujan. Contohnya dalam kajian yang dilakukan oleh Zhao dan Karypis (2002) telah menggunakan teknik pengelompokan hierarki dalam mengelompokkan perkataan-perkataan di dalam sesuatu dokumen. Kajian yang dilakukan oleh Szymkowiak et al (2001) juga menggunakan teknik pengelompokan hierarki dalam menentukan vektor setiap perkataan di dalam sesuatu maklumat *e-mail*. Kajian-kajian yang dijalankan oleh penyelidik sebelum ini hanya menumpukan kepada pengelompokan data berdasarkan set data berbentuk dokumen dan bukan berdasarkan data kaji cuaca. Oleh sebab itu, penyelidikan ini cuba mengaplikasikan teknik pengelompokan data hierarki ke atas data kaji cuaca bagi tujuan peramalan taburan hujan.

Pengelompokan Hierarki dibahagikan kepada dua algoritma iaitu *Agglomerative (bottom-up)* dan *Divisive (top-down)*. Dalam algoritma *Agglomerative*, kelompok dihasilkan melalui pembentukan susunan skema pengelompokan dengan mengurangkan jumlah kelompok pada setiap langkah pengelompokan. Kelompok yang dihasilkan diperolehi

daripada gabungan kelompok-kelompok yang terhampir (sama) kepada satu kelompok. Manakala dalam algoritma *Divisive*, kelompok dijana dengan menghasilkan susunan skima pengelompokan melalui penambahan bilangan kelompok bagi setiap langkah pengelompokan. Kelompok yang dihasilkan adalah dengan memisahkan kelompok kepada dua (Haldiki et al, 2001).

Namun begitu, skop kajian ini hanya tertumpu kepada algoritma *Agglomerative* sebagai kaedah pengelompokan yang akan diguna pakai ke atas set data kajian. Ini memandangkan, bagi setiap kelompok dengan *n* objek pada algoritma *Divisive*, terdapat kemungkinan sebanyak $2^{N-1} - 1$ dua-subset pembahagian perlu dilakukan (Xu dan Wunsch, 2005). Justeru itu, pengelompokan *divisive* tidak banyak digunakan secara praktiknya.

Algoritma *Agglomerative* dibahagikan kepada tiga jenis iaitu algoritma *Single Link*, algoritma *Average Link* dan algoritma *Complete Link*. Setiap algoritma ini mempunyai perbezaan dari segi pengiraan jarak di antara kelompok-kelompok bagi pembentukan sesuatu kelompok. Menurut Lin (1999), algoritma *Agglomerative* dimulakan dengan *n* kelompok dan setiap kelompok mengandungi satu set data bagi atribut yang berkaitan. Selepas itu, setiap kelompok akan digabungkan berdasarkan persamaan ciri di antara kelompok sehingga kelompok dikelompokkan mengikut yang dikehendaki atau sehingga pembentukan satu kelompok. Secara amnya, algoritma *Agglomerative* boleh diringkaskan dengan mengikuti prosidur seperti berikut:

(i) Bermula dengan *n* kelompok dan satu sampel setiap kelompok.

(ii) Cari persamaan antara kelompok dan gabungkan kelompok yang sama.

(iii) Ulang langkah (ii) sehingga kelompok yang dikehendaki atau satu kelompok.

Pengiraan persamaan atau jarak di antara kelompok ditentukan dengan menentukan pasangan kelompok yang sepatutnya dikelompokkan. Pengiraan ini berbeza mengikut jenis algoritma *Agglomerative* yang digunakan. Bagi algoritma *Single Link*, pengiraan jarak di antara kelompok dilakukan dengan mencari jarak yang terdekat di antara sampel dalam satu kelompok dan sampel di dalam kelompok yang lain.

Formula jarak algoritma *Single Link* (Cimiano et al, 2001; Lin, 1999):

$$d(C_i, C_j) = \min_{a \in C_i, b \in C_j} d(a,b) \quad \ldots (1)$$

di mana $d(C_i , C_j)$ merupakan fungsi jarak yang dibincangkan tadi di dalam matriks proximity. Manakala di dalam algoritma *Average Link*, pengiraan jarak di antara kelompok dilakukan dengan menentukan jarak purata di antara sampel di dalam kelompok atau sampel dengan

kelompok lain. Formula jarak algoritma *Average Link* (Cimiano et al, 2001; Lin, 1999) adalah:

$$d(C_i, C_j) = \frac{1}{n_{c_i} n_{c_j}} \sum_{a \in C_i, b \in C_j} d(a,b) \quad \ldots (2)$$

di mana n_{c_i} merupakan jumlah kelompok bagi kelompok C_i. Pengiraan jarak di dalam algoritma *Complete Link* pula diperolehi dengan menentukan jarak yang terpanjang di antara sampel di dalam satu kelompok dan sampel di dalam kelompok yang lain. Formula jarak algoritma *Complete Link* (Cimiano et al, 2001; Lin, 1999) adalah:

$$d(C_i, C_j) = \max_{a \in C_i, b \in C_j} d(a,b) \quad \ldots (3)$$

3.0 Metodologi

Perlaksanaan kajian ini telah dilakukan mengikut beberapa aktiviti. Di antara aktiviti-aktiviti yang terlibat ialah mengumpul dan menganalisis data kaji cuaca, mengira jarak Euclidean, membangunkan atur cara, mengelompokkan set data kaji cuaca, melakukan pengujian kelompok data kaji cuaca, menganalisis keputusan pengujian dan akhir sekali ialah membuat perbandingan keputusan pengujian yang telah dihasilkan.

Bagi menjayakan perlaksanaan kajian ini, sebanyak 500 set data kaji cuaca, bermula dari 1 September 2000 sehingga 21 September 2000 telah dikumpul. Data tersebut telah diperolehi daripada Jabatan Perkhidmatan Kaji cuaca Malaysia. Terdapat lapan pemboleh ubah atau atribut utama data kaji cuaca yang telah diguna pakai di dalam perlaksanaan kajian ini, di antaranya ialah *windvane*, *humidity*, *energy*, *temp*, *tension*, *radiation*, *windspeed* dan *rainfall*.

Selepas data kaji cuaca tersebut dianalisis, pengiraan jarak Euclidean pula dilakukan. Proses ini dijalankan untuk mendapatkan persamaan di antara atribut-atribut data kaji cuaca tersebut untuk tujuan mengelompokkan atribut-atribut yang berkaitan. Formula yang digunakan untuk melakukan pengiraan jarak *Euclidean* ialah :

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \quad \ldots (4)$$

di mana $x = (x_1, x_2, ..., x_n)$ dan $y = (y_1, y_2, ..., y_n)$

Berdasarkan pengiraan jarak Euclidean yang telah dihasilkan, algoritma pengelompokan hierarki bagi *single link*, *average link* dan *complete link* dibangunkan menggunakan bahasa pengaturcaraan C. Dengan menggunakan atur cara yang telah dibangunkan, atribut-atribut data kaji cuaca (kecuali atribut *rainfall*) dikelompokkan kepada beberapa kelompok iaitu kelompok 2, 3, 4, 5 dan 6. Proses ini dilakukan berulang-ulang kali dengan menggunakan beberapa set data yang terdiri daripada 100 set data, 200 set data, 300 set data, 400 set data dan 500 set data kaji cuaca.

Setelah pengelompokan dijalankan ke atas set data kaji cuaca, setiap kelompok yang terhasil diuji dengan menggunakan Pakej NeuNet Pro 2.3. Pengujian ini telah dibahagikan kepada dua kategori, iaitu eksperimen yang melibatkan atribut-atribut dari kelompok yang sama dan juga eksperimen yang melibatkan atribut-atribut dari kelompok yang berlainan. Di dalam pengujian ini, atribut-atribut dari kelompok yang sama dan atribut-atribut dari kelompok yang berlainan digunakan sebagai data input untuk melakukan proses peramalan taburan hujan. Pengujian kelompok ini dilakukan untuk membuat penganalisisan dan perbandingan prestasi peramalan taburan hujan di antara set-set eksperimen yang dijalankan.

Penganalisisan kemudian dilakukan ke atas hasil pengujian tersebut untuk menentukan prestasi teknik pengelompokan menggunakan algoritma pengelompokan hierarki. Penganalisaan ini dilakukan dengan melihat nilai perbezaan di antara nilai sebenar taburan hujan serta nilai ramalan taburan hujan yang dihasilkan. Berdasarkan kepada perbezaan nilai tersebut, nilai ralat min punca kuasa dua (RMS) dan pekali korelasi akan dihasilkan. Sehubungan dengan itu, nilai RMS dan pekali korelasi ini digunakan sebagai perbandingan bagi menentukan keberkesanan teknik pengelompokan hierarki.

4.0 Eksperimen

Di dalam kajian yang dijalankan, proses pengelompokan menggunakan algoritma *Agglomerative* telah dilakukan ke atas 500 set data kaji cuaca. Proses pengelompokan ini telah dibahagikan kepada tiga bahagian utama iaitu pengelompokan menggunakan algoritma *Single Link*, algoritma *Average Link* dan algoritma *Complete Link*. Eksperimen ini dijalankan berulang-ulang kali dengan mengelompokkan 100 set data, 200 set data, 300 set data, 400 set data dan 500 set data kaji cuaca kepada 2 kelompok, 3 kelompok, 4 kelompok, 5 kelompok dan 6 kelompok.

Selepas pengelompokan dijalankan ke atas set-set data tadi, pengujian kelompok pula dilakukan bagi setiap kelompok data kaji cuaca yang telah dihasilkan. Ia melibatkan pengujian peramalan taburan hujan dengan menggunakan atribut-atribut dari kelompok data kaji cuaca yang sama dan juga kelompok data kaji cuaca yang berlainan. Pengujian ini dilakukan dengan menggunakan aplikasi NeuNet Pro 2.3. Keputusan peramalan taburan hujan yang diperolehi digunakan untuk membuat penganalisisan dan perbandingan.

4.1 Algoritma *Single Link*

Berdasarkan eksperimen yang telah dilakukan, Jadual 1 berikut adalah merupakan hasil pengelompokan data kaji cuaca dengan menggunakan algoritma *Single Link*.

Jadual 1 : Hasil pengelompokan data kaji cuaca menggunakan algoritma *Single Link*

Bil. Kelompok	100 Set Data	200 Set Data	300 Set Data	400 Set Data	500 Set Data
2	(a) (b, c,d, e, f, g)	(a, b, d, e, g) (c, f)	(a, b, d, e, g) (c, f)	(a, b, d, e, g) (c, f)	(a, b, d, e, g) (c, f)
3	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)	(a) (b, d, e, g) (c, f)
4	(a) (b, d, e, g) (c) (f)	(a) (b, d, e, g) (c) (f)	(a) (b, d, e, g) (c) (f)	(a) (b, d, e, g) (c) (f)	(a) (b, d, e, g) (c) (f)
5	(a) (b, d, g) (c) (e) (f)	(a) (b, d, g) (c) (e) (f)	(a) (b, d, g) (c) (e) (f)	(a) (b, d, g) (c) (e) (f)	(a) (b, d, g) (c) (e) (f)
6	(a) (b, g) (c) (d) (e) (f)	(a) (b, g) (c) (d) (e) (f)	(a) (b, g) (c) (d) (e) (f)	(a) (b) (c) (d, g) (e) (f)	(a) (b) (c) (d, g) (e) (f)

a – *windvane*, b – *humidity*, c – *energy*, d – *temp*, e – *tension*, f – *radiation*, g – *windspeed*

Setelah kelompok-kelompok data kaji cuaca dikenalpasti, pengujian kelompok pula dilakukan. Proses pengujian ini melibatkan penggunaan atribut-atribut data kaji cuaca (kecuali atribut *rainfall*) sebagai data input untuk melakukan proses peramalan taburan hujan. Pengujian ini telah dibahagikan kepada dua kategori, iaitu pengujian bagi kelompok data kaji

cuaca yang sama dan juga pengujian bagi kelompok data kaji cuaca yang berlainan. Oleh yang demikian, atribut-atribut dari kelompok yang sama dan juga atribut-atribut dari kelompok yang berlainan telah digunakan sebagai data input kepada proses peramalan tersebut. Bagi setiap kategori pengujian, terdapat beberapa eksperimen yang telah dijalankan.

Setiap eksperimen tersebut melibatkan atribut-atribut bagi 2 kelompok, 3 kelompok, 4 kelompok, 5 kelompok dan 6 kelompok data kaji cuaca dan juga menggunakan 100 set data, 200 set data, 300 set data, 400 set data dan 500 set data kaji cuaca. Pengujian bagi setiap kategori kelompok dan set data dilakukan beberapa kali dan nilai purata bagi nilai ralat min punca kuasa dua (RMS) dan pekali korelasi telah diambil dan digunakan untuk tujuan perbandingan. Jadual 2(a) dan 2(b) menunjukkan hasil peramalan taburan hujan yang dilakukan ke atas kelompok atribut data kaji cuaca yang sama dan juga kelompok atribut data kaji cuaca yang berlainan.

Jadual 2(a) : Keputusan peramalan taburan hujan bagi kelompok berlainan

KELOMPOK	**100 Set Data**		**200 Set Data**		**300 Set Data**		**400 Set Data**		**500 Set Data**	
	RMS	CC	RMS	CC	RMS	CC	RMS	CC	RMS	CC
2 Kelompok	0.95	0.07	0.85	0.27	0.88	0.17	0.96	0.12	0.95	0.09
3 Kelompok	0.93	0.15	0.81	0.39	0.85	0.26	0.96	0.16	0.94	0.16
4 Kelompok	0.88	0.37	0.79	0.46	0.85	0.23	0.92	0.33	0.92	0.28
5 Kelompok	0.91	0.22	0.78	0.49	0.78	0.49	0.89	0.25	0.90	0.34
6 Kelompok	0.84	0.38	0.68	0.64	0.74	0.56	0.79	0.67	0.78	0.56

Berdasarkan hasil pengujian bagi kelompok data kaji cuaca berlainan yang telah dilaksanakan, maka penganalisisan terhadap keputusan peramalan taburan hujan telah dibuat. Secara keseluruhannya, didapati kesemua eksperimen yang melibatkan set data 100, 200, 300, 400 dan 500 menunjukkan bahawa hasil atau prestasi peramalan semakin meningkat, selari dengan bilangan atribut-atribut data kaji cuaca yang digunakan di dalam pengujian tersebut. Atau dengan kata lain, semakin banyak bilangan atribut data kaji cuaca yang digunakan sebagai data input kepada proses peramalan taburan hujan tersebut, semakin baik prestasi peramalan yang dihasilkan. Ini dibuktikan oleh nilai RMS dan juga pekali korelasi yang telah dihasilkan di dalam perlaksanaan eksperimen yang berkaitan.

Selain daripada itu, hasil eksperimen juga menunjukkan keputusan yang dihasilkan oleh eksperimen yang melibatkan set data kaji cuaca 200 memberikan keputusan peramalan yang terbaik berbanding set data kaji cuaca yang lain bagi semua kelompok 2, 3, 4, 5 dan 6

yang dijalankan. Keadaan ini mungkin juga disebabkan oleh kewujudan bilangan data melampau yang banyak di dalam set data 300, 400 dan 500.

Jadual 2(b) : Keputusan peramalan taburan hujan bagi kelompok sama

KELOMPOK	100 Set Data		200 Set Data		300 Set Data		400 Set Data		500 Set Data	
	RMS	CC	RMS	CC	RMS	CC	RMS	CC	RMS	CC
2 Kelompok	0.87	0.29	0.82	0.29	0.77	0.36	0.88	0.30	0.90	0.28
3 Kelompok	0.94	0.12	0.89	0.08	0.84	0.19	0.85	0.21	0.95	0.13
4 Kelompok	0.70	0.07	0.67	0.06	0.63	0.14	0.77	0.16	0.69	0.14
5 Kelompok	0.95	0.03	0.86	0.15	0.86	0.14	1.03	0.07	0.96	0.08
6 Kelompok	0.95	0.03	0.90	0.06	0.88	0.10	0.93	0.03	0.80	0.06

Manakala bagi eksperimen yang melibatkan penggunaan atribut-atribut dari kelompok data kaji cuaca yang sama pula menunjukkan keputusan peramalan taburan hujan yang berbeza. Berdasarkan kepada keputusan pada Jadual 2(b) di atas, maka dapatlah dirumuskan bahawa keputusan peramalan yang dihasilkan oleh 4 kelompok memberikan prestasi yang terbaik berbanding dengan empat eksperimen yang lain. Prestasi ini boleh dilihat pada nilai RMS dan juga nilai pekali korelasi yang dihasilkan oleh eksperimen yang melibatkan penggunaaan atribut-atribut data kaji cuaca dari 4 kelompok di dalam kesemua eskperimen sama ada yang melibatkan 100 data set, 200 data set, 300 data set, 400 data set, mahupun 500 data set. Di samping itu, keputusan peramalan taburan hujan yang ditunjukkan oleh eksperimen yang melibatkan 300 set data kaji cuaca juga memberikan keputusan prestasi peramalan yang lebih baik jika dibandingkan dengan prestasi peramalan taburan hujan yang melibatkan set data yang lain.

Oleh yang demikian, secara keseluruhannya bolehlah dirumuskan bahawa keputusan peramalan taburan hujan yang dihasilkan oleh eskperimen yang melibatkan penggunaan atribut dari kelompok data kaji cuaca yang berlainan menunjukkan prestasi peramalan taburan hujan yang lebih baik berbanding dengan prestasi peramalan taburan hujan yang dihasilkan oleh eksperimen yang melibatkan penggunaan atribut dari kelompok data kaji cuaca yang sama. Ini dibuktikan oleh keputusan peramalan taburan hujan yang dihasilkan di dalam kelima-lima eksperimen yang melibatkan bilangan set data kaji cuaca yang berbeza.

Selain daripada itu, keputusan peramalan bagi eksperimen kelompok berlainan menunjukkan semakin banyak bilangan atribut data kaji cuaca yang digunakan di dalam pengujian tersebut, semakin tinggi prestasi peramalan yang dihasilkan. Ini dapat dilihat pada nilai RMS yang semakin berkurang dan juga nilai pekali korelasi yang semakin meningkat.

5.2 Algoritma *Average Link*

Proses pengelompokan telah dilaksanakan ke atas set data kaji cuaca dengan menggunakan algoritma *Average Link*. Jadual 3 menunjukkan hasil pengelompokan data kaji cuaca tersebut.

Jadual 3 : Hasil pengelompokan data kaji cuaca menggunakan algoritma *Average Link*

Bil. Kelompok	100 Set Data	200 Set Data	300 Set Data	400 Set Data	500 Set Data
2	(a,c,d,e,f,g) (b)	(a,c,d,e,f,g) (b)	(a,c,d,e,f,g) (b)	(a,c,d,e,f,g) (b)	(a,c,d,e,f,g) (b)
3	(a) (b) (c,d,e,f,g)	(a) (b,c, d, e, f) (g)	(a, c, d, e, g) (b) (f)	(a, c, d, e, g) (b) (f)	(a, c, d, e, g) (b) (f)
4	(a) (b) (c,d,f,g) (e)	(a) (b,c,d, f) (e) (g)	(a, c, d, e) (b) (f) (g)	(a, c, d, e) (b) (f) (g)	(a, c, d, e) (b) (f) (g)
5	(a) (b) (c,f,g) (d) (e)	(a) (b) (c, d , f) (e) (g)	(a, e, d) (b) (e) (f) (g)	(a, e, d) (b) (e) (f) (g)	(a, e, d) (b) (e) (f) (g)
6	(a) (b) (c,g) (d) (e) (f)	(a) (b) (c, d) (e) (f) (g)	(a) (b) (c, d) (e) (f) (g)	(a) (b) (c, d) (e) (f) (g)	(a) (b) (c, d) (e) (f) (g)

a – *windvane*, b – *humidity*, c – *energy*, d – *temp*, e – *tension*, f – *radiation*, g – *windspeed*

Berdasarkan kelompok-kelompok data kaji cuaca yang telah dihasilkan dengan menggunakan algoritma *Average Link*, pengujian kelompok dilaksanakan dengan menjalankan beberapa eksperimen. Di antaranya melibatkan penggunaan atribut dari kelompok yang sama dan juga dari kelompok yang berlainan. Di dalam perlaksanaan eksperimen tersebut, atribut data kaji cuaca (atribut selain *rainfall*) telah digunakan sebagai data input kepada proses pengujian, iaitu proses peramalan taburan hujan. Keputusan yang dihasilkan di dalam setiap eskperimen tersebut, akan dicatatkan dan nilai purata bagi nilai ralat min punca kuasa dua (RMS) dan juga nilai pekali korelasi digunakan untuk perbandingan keputusan. Jadual 4(a) dan 4(b) menunjukkan hasil purata bagi nilai RMS dan pekali korelasi yang dihasilkan oleh setiap eksperimen yang dijalankan.

Jadual 4(a) : Keputusan peramalan taburan hujan bagi kelompok berlainan

KELOMPOK	100 Set Data		200 Set Data		300 Set Data		400 Set Data		500 Set Data	
	RMS	CC	RMS	CC	RMS	CC	RMS	CC	RMS	CC
2 Kelompok	0.94	0.09	0.87	0.24	0.74	0.15	1.09	0.14	0.96	0.10
3 Kelompok	0.94	0.13	0.83	0.35	0.86	0.23	0.94	0.26	0.78	0.15
4 Kelompok	0.95	0.09	0.85	0.30	0.74	0.54	0.86	0.62	0.81	0.53
5 Kelompok	0.89	0.33	0.70	0.62	0.68	0.64	0.77	0.71	0.74	0.63
6 Kelompok	0.94	0.17	0.58	0.76	0.75	0.55	0.76	0.73	0.76	0.61

Jadual 4(a) di atas menunjukkan keputusan peramalan taburan hujan bagi kelompok berlainan. Bagi eksperimen yang melibatkan penggunaan atribut dari kelompok data kaji cuaca yang berlainan sebagai data input kepada pengujian, didapati keputusan peramalan taburan hujan yang dihasilkan memberikan prestasi peramalan yang semakin baik atau meningkat selari dengan pertambahan bilangan atribut data kaji cuaca yang digunakan. Ini ditunjukkan oleh kesemua eskperimen yang melibatkan penggunaan data set (kecuali 100 set data menunjukkan prestasi yang tidak sekata). Jadual 4(a) juga menunjukkan prestasi peramalan taburan hujan yang dihasilkan di dalam eksperimen yang melibatkan penggunaan 400 set data memberikan prestasi yang terbaik berbanding dengan eksperimen yang lain.

Jadual 4(b) : Keputusan peramalan taburan hujan bagi kelompok sama

KELOMPOK	100 Set Data		200 Set Data		300 Set Data		400 Set Data		500 Set Data	
	RMS	CC	RMS	CC	RMS	CC	RMS	CC	RMS	CC
2 Kelompok	0.92	0.04	0.74	0.43	0.81	0.29	0.97	0.34	0.84	0.34
3 Kelompok	0.92	0.16	0.88	0.13	0.84	0.18	0.55	0.67	0.95	0.13
4 Kelompok	0.94	0.10	0.89	0.11	0.88	0.11	1.11	0.05	0.97	0.04
5 Kelompok	0.95	0.05	0.87	0.15	0.89	0.09	1.11	0.24	0.97	0.01
6 Kelompok	0.95	0.03	0.89	0.08	0.89	0.07	1.11	0.06	0.97	0.01

Manakala Jadual 4(b) pula menunjukkan hasil peramalan taburan hujan yang dilakukan ke atas kelompok atribut data kaji cuaca yang sama. Berdasarkan kepada keputusan peramalan taburan hujan yang ditunjukkan di dalam jadual di atas, didapati semakin banyak bilangan atribut data kaji cuaca yang digunakan sebagai data input kepada proses pengujian, semakin menurun prestasi peramalan taburan hujan yang dihasilkan. Prestasi ini ditunjukkan oleh nilai RMS yang semakin tinggi dan juga nilai pekali korelasi

yang semakin menghampiri 0 di dalam kesemua eksperimen yang melibatkan 100 set data, 200 set data, 300 set data, 400 set data dan 500 set data kaji cuaca.

Oleh yang demikian, dapatlah dirumuskan bahawa dengan menggunakan algoritma *Average Link* untuk mengelompokkan data kaji cuaca, didapati keputusan peramalan taburan hujan yang dihasilkan oleh kelompok yang berlainan memberikan prestasi yang lebih baik berbanding dengan keputusan peramalan taburan hujan yang dihasilkan oleh kelompok yang sama.

5.3 Algoritma *Complete Link*

Jadual 5 menunjukkan hasil pengelompokan data kaji cuaca yang dilakukan dengan menggunakan algoritma *Complete Link*. Berdasarkan hasil pengelompokan ini, beberapa eksperimen telah dilaksanakan untuk melihat prestasi peramalan taburan hujan yang dihasilkan. Eksperimen ini dijalankan dengan melibatkan penggunaan atribut-atribut data kaji cuaca dari kelompok yang sama dan juga kelompok yang berlainan.

Jadual 5 : Hasil pengelompokan data kaji cuaca menggunakan algoritma *Complete Link*

Bil. Kelompok	100 Set Data	200 Set Data	300 Set Data	400 Set Data	500 Set Data
2	(a, d, f, g) (b, c, e)	(a, d, f, g) (b, c, e)	(a, c, d) (b, e, f, g)	(a, c, d) (b, e, f, g)	(a, b, c, d) (e, f, g)
3	(a, d, f, g) (b) (c, e)	(a, d, f, g) (b) (c, e)	(a, c, d) (b) (e, f, g)	(a, c, d) (b) (e, f, g)	(a, c, d) (b) (e, f, g)
4	(a, f, g) (b) (c, e) (d)	(a, f, g) (b) (c, e) (d)	(a, c) (b) (d) (e, f, g)	(a, c, d) (b) (e) (f, g)	(a, c) (b) (d) (e, f, g)
5	(a, f, g) (b) (c) (d) (e)	(a, f, g) (b) (c) (d) (e)	(a, c) (b) (d) (e, f) (g)	(a, c) (b) (d) (e) (f, g)	(a, c) (b) (d) (e, f) (g)
6	(a, f) (b) (c) (d) (e) (g)	(a, f) (b) (c) (d) (e) (g)	(a) (b) (c) (d) (e, f) (g)	(a) (b) (c) (d) (e) (f, g)	(a) (b) (c) (d) (e, f) (g)

a – *windvane*, b – *humidity*, c – *energy*, d – *temp*, e – *tension*, f – *radiation*, g - *windspeed*

Jadual 6(a) menunjukkan keputusan peramalan taburan hujan yang melibatkan penggunaan atribut data kaji cuaca dari kelompok berlainan. Manakala 6(b) pula menunjukkan keputusan peramalan taburan hujan yang menggunakan atribut data kaji cuaca dari kelompok sama sebagai data input kepada eksperimen yang telah dijalankan.

Jadual 6(a) : Keputusan peramalan taburan hujan bagi kelompok berlainan

KELOMPOK	100 Set Data		200 Set Data		300 Set Data		400 Set Data		500 Set Data	
	RMS	CC	RMS	CC	RMS	CC	RMS	CC	RMS	CC
2 Kelompok	0.94	0.09	0.87	0.22	0.88	0.15	1.06	0.10	0.96	0.11
3 Kelompok	0.95	0.08	0.83	0.32	0.84	0.28	0.95	0.34	0.94	0.19
4 Kelompok	0.71	0.27	0.73	0.52	0.76	0.48	0.90	0.55	0.91	0.32
5 Kelompok	0.74	0.16	0.71	0.57	0.60	0.71	0.89	0.59	0.76	0.61
6 Kelompok	0.72	0.25	0.71	0.60	0.56	0.75	0.76	0.73	0.76	0.61

Bagi eksperimen yang melibatkan penggunaan atribut dari kelompok yang sama, didapati keputusan peramalan taburan hujan yang dihasilkan menunjukkan semakin banyak bilangan atribut data kaji cuaca yang digunakan sebagai data input kepada proses peramalan, semakin meningkat prestasi peramalan taburan hujan yang dihasilkan. Ini ditunjukkan oleh kesemua eksperimen yang melibatkan 100 set data, 200 set data, 300 set data, 400 set data dan 500 set data, di mana nilai RMS yang dihasilkan semakin menurun manakala nilai pekali korelasinya semakin meningkat. Selain daripada itu, didapati eksperimen yang menggunakan 200 set data, 300 set data dan 400 set data menunjukkan prestasi peramalan yang lebih baik berbanding yang lain.

Jadual 6(b) : Keputusan peramalan taburan hujan bagi kelompok sama

KELOMPOK	100 Set Data		200 Set Data		300 Set Data		400 Set Data		500 Set Data	
	RMS	CC	RMS	CC	RMS	CC	RMS	CC	RMS	CC
2 Kelompok	0.95	0.08	0.95	0.09	0.83	0.35	0.91	0.49	0.87	0.41
3 Kelompok	0.94	0.11	0.80	0.33	0.84	0.30	0.90	0.26	0.92	0.19
4 Kelompok	0.95	0.06	0.87	0.22	0.87	0.16	0.95	0.18	0.93	0.14
5 Kelompok	0.95	0.05	0.87	0.17	0.88	0.12	0.95	0.10	0.98	0.04
6 Kelompok	0.95	0.04	0.89	0.10	0.88	0.16	0.95	0.09	0.98	0.03

Berdasarkan keputusan peramalan yang dipaparkan pada Jadual 6(b) di atas, didapati prestasi peramalan yang dihasilkan semakin menurun selari dengan peningkatan bilangan atribut data kaji cuaca yang digunakan di dalam proses peramalan taburan hujan tersebut. Ini ditunjukkan oleh kesemua eksperimen yang dijalankan, yang melibatkan penggunaan 100 set data, 200 set data, 300 set data, 400 set data dan 500 set data.

Oleh yang demikian, maka bolehlah dirumuskan bahawa dengan mengelompokkan data kaji cuaca menggunakan algoritma *Complete Link*, didapati prestasi peramalan taburan hujan yang dihasilkan dengan menggunakan atribut data kaji cuaca dari kelompok berlainan adalah lebih baik jika dibandingkan dengan prestasi peramalan taburan hujan yang dihasilkan dengan menggunakan atribut data kaji cuaca dari kelompok sama. Ini dibuktikan oleh nilai RMS dan pekali korelasinya yang ditunjukkan pada Jadual 6(a) dan Jadual 6(b).

6.0 Kesimpulan

Berdasarkan kepada keputusan eksperimen yang telah dijalankan, maka secara keseluruhannya bolehlah dirumuskan bahawa prestasi peramalan taburan hujan yang dihasilkan oleh eksperimen yang melibatkan penggunaan atribut-atribut data kaji cuaca dari kelompok yang berlainan adalah lebih baik jika dibandingkan dengan prestasi peramalan taburan hujan yang menggunakan atribut data kaji cuaca dari kelompok yang sama. Ini ditunjukkan oleh nilai RMS eksperimennya yang semakin menghampiri 0 dan juga nilai pekali korelasinya yang semakin menghampiri 1 di dalam kesemua eksperimen yang melibatkan ketiga-tiga teknik pengelompokan iaitu algoritma *Single Link*, *Average Link* dan juga *Complete Link*.

Walau bagaimanapun, didapati algoritma *Complete Link* memberikan prestasi peramalan taburan hujan yang lebih baik berbanding dua algoritma yang lain, diikuti oleh algoritma *Single Link* dan akhir sekali *Average Link*. Sehubungan dengan itu, kajian ini boleh diteruskan dengan membuat perbandingan di antara teknik pengelompokan Hierarki *Agglomorative* dengan teknik-teknik pengelompokan yang lain, contohnya teknik pengelompokan Hierarki *Divisive* (seperti *Divisive Analysis* dan *Monothetic Analysis*) dan teknik pengelompokan *Partitional* (seperti *Square Error, Graph Theoretic, Mixture Resolving* dan *Mode Seeking*).

Rujukan

Chen, T., and Takagi, M. (1993). Rainfall prediction of geostationary Meteorological satellite images using artificial neural network. *International Geoscience and Remote Sensing Symposium*. 3:1247-1249.

Cimiano, P., Hotho, A. and Staab, S. (2001). Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text.

Haldiki, M., Batistakis, Y., and Vazirgiannis, M. (2001). Clustering algorithms and validity measures. Tutorial paper, *Proceedings of SSDBM Conference*.3-22.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data Clustering : A review. *ACM Computing Surveys (CSUR)*. 31(3):264-323.

Lin, H. (1999). Survey and implementation of clustering algorithms. Theses. Hsinchu, Taiwan, Republic of China.

Liu, J. N. K., and Lee, R. S. T. (1999). Rainfall forecasting from multiple point sources using neural networks. *In Proceedings of the 1999 IEEE International Conference on Systems, Man, and Cybernetics (SMC '99)*. 3:429-434.

Malaysia Meteorological Services (2004).
[Online] Available : http://www.kjc.gov.my/

McCullagh, J., Bluff, K., and Ebert, E. (1995). A Neural network model for rainfall estimation. *The Second New Zealand International Two Stream Conference on Artificial Neural Networks and Expert Systems.* 389-392.

McCullagh, J., Bluff, K., and Hendtlass, T. (1999). Evolving expert neural networks for meteorological rainfall estimations. Proceedings of the *International Conference on Neural Information Processing and Intelligent Information Systems IEEE (ICONIP '99)*. 2:585-590.

Ochiai, K., Suzuki, H., Shinozawa, K., Fujii, M. and Sonehara, N. (1995). Snowfall and rainfall forecasting from weather radar images with artificial neural networks. *Proceedings of IEEE International Conferenc*e. 2:1182-1187.

Szymkowiak, A., Larsen, J. and Hansen, L. K. (2001). Hierarchical clustering for data mining.

Ultsch, A., and Guimareas, G. (1996). Classification and prediction of hail using self-organizing neural networks. *In Proceedings of the International Conference on Neural Networks ICNN '96*. 1622-1627.

V. Diyankov, Vladimir A. Lykov and Serge A. Terekhoff (1992). Artificial neural networks in weather forecasting.

Xu, R. and Wunsch, D. (2005). Survey of Clustering Algorithms. *IEE Transactions on Neural Networks.* Vol 16. 645-678.

Zhao, Y. and Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. *(CIKM'02)*. 515-524.

PENGHARGAAN

Dengan Nama Allah Yang Maha Pemurah Lagi Maha Penyayang

Alhamdulillah, saya bersyukur kehadrat Allah S.W.T. yang telah melimpahkan rahmat dan mengurniakan keyakinan, kekuatan dan semangat serta mengizinkan saya menyiapkan laporan projek II ini.

Setinggi-tinggi penghargaan saya tujukan kepada rakan-rakan penyelidik Rozilawati, Pm Dr Mohd Noor dan Aryati yang sama-sama membantu di dalam menjayakan kajian ini. Tidak dilupakan juga, En. Azhar bin Ishak, En. Santira dan En. Junaidi dari Jabatan Perkhidmatan Kajicuaca Malaysia serta Dr. Khalid dari stesen penyelidikan, Lembaga Minyak Sawit Malaysia, Kluang di atas segala penerangan dan bantuan anda.

ABSTRAK

Perolehan maklumat yang berguna daripada timbunan data kajicuaca menjadi sukar ekoran daripada pertambahan jumlah data yang disimpan di JPKM. Ini kerana parameter dan jumlah data kajicuaca tersebut semakin meningkat dari masa ke semasa. Jumlah data yang besar ini telah menyukarkan kerja-kerja penganalisaan data kajicuaca bagi tujuan peramalan taburan hujan. Di dalam proses peramalan taburan hujan, adalah tidak munasabah untuk menggunakan kesemua parameter kajicuaca untuk melakukan peramalan. Oleh yang demikian, salah satu cara untuk mengenalpasti parameter manakah yang memberikan pengaruh kepada ketepatan atau prestasi hasil peramalan taburan hujan ialah dengan melakukan pengelompokan ke atas data kajicuaca tersebut. Kajian ini bertujuan untuk mengkaji dan membuat perbandingan di antara dua teknik pengelompokan iaitu kaedah *partitional* dan *hierarchical* untuk melaksanakan pengelompokan data kajicuaca bagi tujuan peramalan taburan hujan. Hasil kajian ini mendapati bahawa pengelompokan *partitional* adalah lebih sesuai untuk digunakan di dalam pengelompokan data kajicuaca berbanding dengan pengelompokan *hierarchical.* Selain daripada itu, penggunaan atribut data kajicuaca yang berada di dalam kelompok yang berlainan memberikan prestasi peramalan yang lebih baik daripada penggunaan atribut data kajicuaca yang berada di dalam kelompok yang sama.

ABSTRACT

Gaining useful information from a stack of weather forecast data is a quiet difficult task as the data keeps on increasing within the JKPM data reservoir. This is due to the increment of parameters and total weather forecast data from time to time. The vast total of data causes difficulties in the process of data analysis for the purpose of rain distribution forecasting. It is not reasonable to use all of the parameters in this forecasting process. As a result, one possible way to recognize which of the parameters that give influence on the accurateness and the performance of the rain distribution forecasting results is by clustering the data. This study is done to investigate and to make comparisons between two clustering techniques, which are the partition clustering technique and the hierarchical method to cluster the data in order to forecast the rain distribution. The result of this study shows that the partition clustering technique is more suitable in terms of weather forecasting data compared to the statistical method. Besides that, the usage of weather forecast data attribute, which is from a different cluster, results in a better performance of forecasting compared to by using the data from the same cluster.

Rujukan

Al-Harbi, S. H., Rayward-Smith, V. J. (2003). The use of a supervised k-means algorithm on real-valued data with applications in health. *IEA/AIE*. 575-581.

Bdanyopadhyay, S., dan Maulik, U. (2002). An evolutionary technique based on k-means algorithm for optimal clustering in R. *Information Science*. 146:221-237.

Chen, T., dan Takagi, M. (1993). Rainfall prediction of geostationary Meteorological satellite images using artificial neural network. *International Geoscience dan Remote Sensing Symposium*. 3:1247-1249.

Cheung, Y. (2003). K*-means : A new generalized k-means clustering algorithm. *Pattern Recognition Letters*. 24(15):2883-2893.

Chen, T., and Takagi, M. (1993). Rainfall prediction of geostationary Meteorological satellite images using artificial neural network. *International Geoscience and Remote Sensing Symposium*. 3:1247-1249.

Cimiano, P., Hotho, A. and Staab, S. (2001). Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text.

Doherty, K.A.J., Adams, R.G., Davey, N. (2001). Non-Euclidean Norms dan Data Normalization.

Dunham, M.H. (2002). *Data Mining Introductory dan Advanced Topics*. Upper Saddle River, New Jersey.

Fred, A., dan Jain, A. K. (2002). Evidence accumulation clustering based on the k-means algorithm. *Structural, Syntactic, dan Statistical Pattern Recognition, LNCS*. 2396:442-451.

Ganguly, A. R. (2002). A hybrid approach to improving rainfall forecasts. *Computing in Science dan Enfineering.* 4(4):14-21.

Haldiki, M., Batistakis, Y., dan Vazirgiannis, M. (2001). Clustering algorithms dan validity measures. Tutorial paper, *Proceedings of SSDBM Conference*.3-22.

Hartigan, J. A. (1975). *Clustering Algorithms*. Canada. 1-14.

Hirano, S., Sun, X., dan Tsumoto, S. (2004). Comparison of Clustering Methods for Clinical Database. *Information Science.* 159(2):155-165.

Jain, A. K., Murty, M. N., dan Flynn, P. J. (1999). Data Clustering : A review. *ACM Computing Surveys (CSUR)*. 31(3):264-323.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C., Silverman, R., dan Wu, A. Y. (2001). The analysis of a simple k-means clustering algorithm. *Symposium on Computational Geometry*. 100-109.

Kim, B. J., Kripalani, R. H., Oh, J. H., dan Moon, S. E. (2002). Summer monsoon rainfall patterns over South Korea dan associated circulation features. *Theoretical dan Applied Climatology.* 72:65-74.

Kim, J., dan Miller, N. L. (1996). Simulating Winds dan Floods : Regional weather-river prediction dan regional climate research. *IEEE Potentials*. 15(4):17-20.

Kulkarni, A., dan Kripalani, R. H. (1998). Rainfall patterns over India : Classification with Fuzzy C-means method. *Theoretical dan Applied Climatology.* 59:137-146.

Lin, H. (1999). Survey dan implementation of clustering algorithms. Theses. Hsinchu, Taiwan, Republic of China.

Liu, J. N. K., dan Lee, R. S. T. (1999). Rainfall forecasting from multiple point sources using neural networks. *In Proceedings of the 1999 IEEE International Conference on Systems, Man, dan Cybernetics (SMC '99)*. 3:429-434.

Malaysia Meteorological Services (2004).
[Online] Available : http://www.kjc.gov.my/

McCullagh, J., Bluff, K., and Ebert, E. (1995). A Neural network model for rainfall estimation. *The Second New Zealand International Two Stream Conference on Artificial Neural Networks and Expert Systems.* 389-392.

McCullagh, J., Bluff, K., and Hendtlass, T. (1999). Evolving expert neural networks for meteorological rainfall estimations. Proceedings of the *International Conference on Neural Information Processing and Intelligent Information Systems IEEE (ICONIP '99).* 2:585-590.

Ochiai, K., Suzuki, H., Shinozawa, K., Fujii, M. and Sonehara, N. (1995). Snowfall and rainfall forecasting from weather radar images with artificial neural networks. *Proceedings of IEEE International Conferenc*e. 2:1182-1187.

Pena, J. M., Lozana, J. A., dan Larranaga, P. (1999). An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recognition Letters*. 20(10):1027-1040.

Phillips, S. J. (2002). Accelaration of k-means dan related clustering algorithm. *Revised Papers from the 4th International Workshop on Algorithm Engineering dan Experiments*. 166-177.

Smith, K. A., dan Ng, A. (2003). Web page clustering using a self-organizing map of user navigation patterns. *Decisions Support Systems*. 35:245-256.

Szymkowiak, A., Larsen, J. and Hansen, L. K. (2001). Hierarchical clustering for data mining.

Tarsitano, A. (2003). A computational study of several relocation methods for k-means algorithms. *Pattern Recognition Letters*. 36(12):2955-2966.

Ultsch, A., dan Guimareas, G. (1996). Classification dan prediction of hail using self-organizing neural networks. *In Proceedings of the International Conference on Neural Networks ICNN '96*. 1622-1627.

Ultsch, A., and Guimareas, G. (1996). Classification and prediction of hail using self-organizing neural networks. *In Proceedings of the International Conference on Neural Networks ICNN '96*. 1622-1627.

V. Diyankov, Vladimir A. Lykov and Serge A. Terekhoff (1992). Artificial neural networks in weather forecasting.

Wan, S. J., Wong, S. K. M., dan Prusinkiewicz, P. (1988). An algorithmfor multidimensional data clustering. *ACM Transactions on Mathematical Software*. 14(4):153-162.

Xu, R. and Wunsch, D. (2005). Survey of Clustering Algorithms. *IEE Transactions on Neural Networks*. Vol 16. 645-678.

Zait, M., Messatfa, H. (1997). A Comparative Study of Clustering Methods. *Future Generation Computer Systems*, 13:149-159.

Zhao, Y. and Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. *(CIKM'02)*. 515-524.