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)].

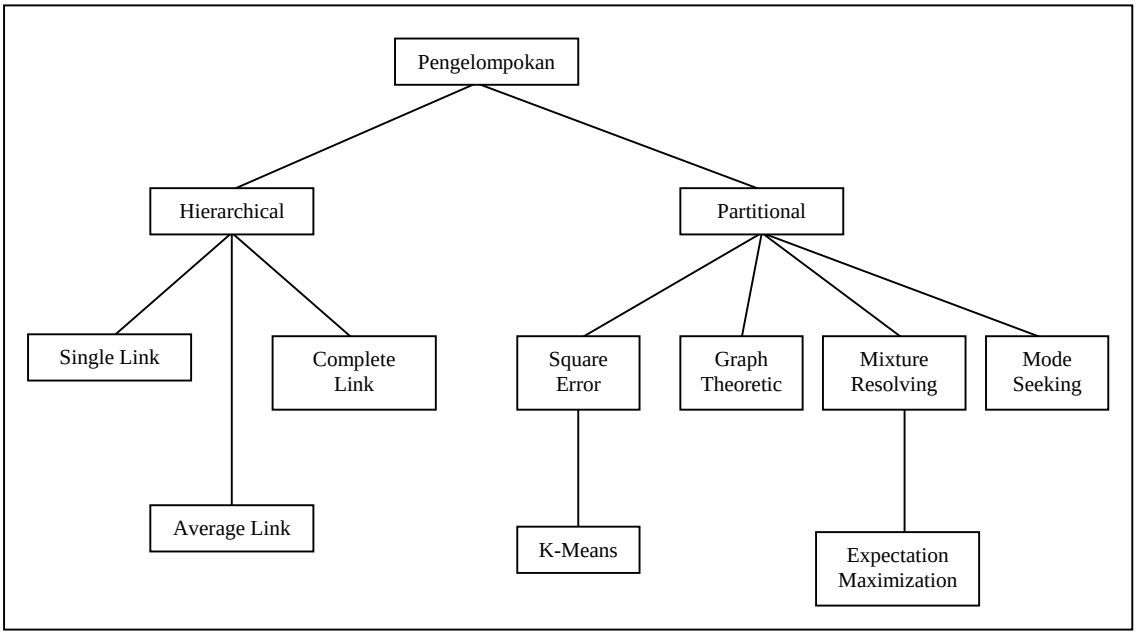
Hujan merupakan salah satu elemen yang rumit serta kompleks untuk diramal kerana ianya mempunyai kepelbagaian pembolehubah atau parameter yang wujud serta mempunyai kaitan yang kompleks antara satu sama lain [Chen dan Takagi(1993); Ultsch dan Guimareas(1996)]. Menurut JPKM, terdapat pelbagai parameter 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