BAB II
TINJAUAN PUSTAKA

A. Tinjauan Pustaka

Penambangan data atau yang sering disebut dengan istilah penambangan data adalah suatu proses iterasi yang dilakukan untuk menemukan suatu pengetahuan yang terdapat dalam kumpulan data tersebut. Applikasi penambangan data ini banyak digunakan oleh beberapa peneliti untuk melakukan penelitian terhadap suatu permasalahan, diantaranya Enrique, et. al (2009) menggunakan *Assosicion Rule Mining* dan *Collaborative Filtering* untuk memberikan rekomendasi materi dan pelajaran apa yang cocok untuk siswa yang akan menggunakan *Learning Management Systems* yang dilengkapi dengan beberapa fitur, seperti forum, diskusi, dan lainnya. Menurut Fabi, et. al (2011), terdapat fase prediksi yang ada dalam metode *Associative Classification*, yaitu dengan metode *Dominant Class Label*, *Highest Average Confidence per Class*, dan *Full Match Rule*. Dengan adanya fase prediksi pada metode *Association Rule Mining* dapat membantu memberikan rekomendasi yang lebih bagus dibandingkan dengan penggunaan teori probabilitas ataupun *decision tree*.

Penggunaan penambangan data tidak hanya terbatas pada bidang teknologi informasi. Penambangan data telah digunakan di bidang ekonomi (Baicoianu dan Dumitrescu, 2010). Beberapa applikasi penambangan data di bidang ekonomi adalah melakukan prediksi kesulitan ekonomi (Shiri, Amini, dan Raftar, 2012), melakukan prediksi terhadap kebangkrutan suatu perusahaan (Shah dan Murtaza, 2000), menentukan resiko kredit dengan menggabungkan penambangan data dan jaringan syaraf tiruan (Pacelli dan Azzollini, 2011), memprediksi nilai tukar dollar selandia baru terhadap dollar amerika (Vojinovic, et. al., 2001), dan beberapa applikasi lainnya. Baicoianu dan Dumitrescu (2010) menyimpulkan bahwa dengan menggunakan pendekatan penambangan data, analisa data ekonomi yang sangat besar dapat menjadi lebih efisien.

Metode *Association Rule Mining* juga digunakan oleh Buddhakulsomsiri, et al (2006) dalam meminimalisasi kerugian akibat besarnya biaya jaminan untuk produk – produk otomotif dan Stergios (2006) untuk mendeteksi kemiripan yang terjadi dalam satu kelompok. Stergios (2006) menggabungkan metode *Fuzzy Association Rules* dengan Klasterisasi dan menggunakan pendekatan *Self Organized Map* (SOM). Selain metode *Association Rule Mining*, metode Bayes juga digunakan oleh Lamm – Tenant, Starks, dan Stokes (1992) untuk melalukan perkiraan terhadap persentase kerugian yang mungkin ditanggung oleh investor. Dengan adanya sistem ini dapat memberikan suatu rekomendasi untuk meminimalisasi kerugian yang ada. Selain itu, terdapat metode Fuzzy untuk membantu dalam dynamic reasoning (Song dan Lee, 2002). Dengan adanya metode Fuzzy membantu sistem menjadi lebih dinamis (Song dan Lee, 2002; Jain, Wadhwa, dan Deshmukh, 2007). Hal ini juga digunakan oleh Jain, Wadhwa, dan Deshmukh (2007) untuk melakukan pemilihan *supplier*. Metode penambangan data lainnya adalah teknik klasterisasi.

Teknik klasterisasi adalah salah satu teknik di dalam penambangan data, dimana data yang ada di kelompokkan berdasarkan kesamaan (*similarity*) dan ketidaksamaan (*dissimilarity*) yang ada (Andayani, 2007). Teknik Klasterisasi dapat digunakan dalam beberapa applikasi untuk mengelompokan suatu permasalahan berdasarkan trend tertentu, diantaranya untuk mengelompokkan data pelanggan (Romdhane, Fadhel, dan Ayeb; 2009), pengelompokan terhadap volatilias harga perumahan di Amerika Serikat (Miles; 2008), klasifikasi terhadap lalu lintas jaringan (Yingqiu, Li Wei, dan Yunchin; 2007), penentuan nilai Ujian (Suprihatin; 2011), melakukan deteksi terhadap *financial fraud* (Sabau, 2012), melakukan deteksi terhadap kesalahan dalam pemberian obat (Chen, et. al; 2011), penggunaan metode K-medoids untuk penerapan klasterisasi pada *time series data* (Liao, Ting, dan Chang, 2006) dan penggunaan metode klasterisasi untuk melakukan prediksi terhadap hasil akademik siswa (Oyelade, Oladipupo, dan Obagbuwa; 2010).

Teknik klasterisasi juga bermanfaat untuk melakukan pengelompokan terhadap pasar perumahan yang berada di Turki (Hepsen dan Vatansever, 2012).

Dalam pengelompokan ini dihasilkan tiga klaster untuk pengelompokan pasar perumahan di Turki, yaitu: (1) distrik yang memiliki penghasilan sewa rendah pada periode 2007 – 2011. Kelompok ini terdiri dari 29 area; (2) distrik yang memiliki penghasilan sewa lebih tinggi daripada klaster pertama. Kelompok ini terdiri dari 34 area; (3) distrik yang memiliki pengasilan sewa tertinggi dibandingkan dua klaster lainnya.

Blogojevic (2011) melakukan pengelompokan terhadap kebiasaan siswa dalam menggunakan *e-Learning*. Di dalam kesimpulannya, penulis mengelompokkan kebiasaan siswa dalam menggunakan tools e-Learning yang dimiliki oleh sebuah institusi pendidikan dalam membantu proses belajar mengajar di sebuah institusi pendidikan. Berbeda dengan Blogojevic (2011), Wang, et. al (2004) menggunakan klasterisasi untuk melakukan analisa terhadap profile pelajar dalam menggunakan web untuk mendukung proses belajar mengajar di Universitas Nasional Sun Yat – San. Di dapatkan empat jenis pelajar, yaitu : (1) pelajar agresif yang menggunakan web sebagai metode pembelajaran; (2) pelajar aktif yang sedikit berbeda dengan pelajar agresif; (3) pelajar pasif yang menggunakan waktu lebih sedikit dibandingkan dengan pelajar aktif; (4) pembaca non aktif yang tidak pernah menggunakan web sebagai tools pembelajaran. Sedangkan Umran dan Abidin (2009), menggunakan metode k-Means klasterisasi dan *Singular Value Decomposition* (SVD) untuk melakukan pengelompokan dokumen yang ada.

Dari beberapa penelitian tersebut, para penulis menyimpulkan bahwa teknik yang digunakan mampu memberikan suatu hasil klaster yang cukup memadai. Bahkan Oyelade, Oladipupo, dan Obagbuwa (2010) melaporkan bahwa hasil penelitian dapat digunakan untuk memberikan masukan kepada para konsultan akademik dalam merencanakan rencana belajar bagi siswa tersebut. Selain itu, Andayani (2007) mengatakan bahwa klasterisasi dengan menggunakan metode K-Means dapat membantu dalam pembentukan *Knowledge Discovery* dalam suatu basis data. Hal ini juga ditekankan oleh Dragut (2012) yang menyatakan bahwa dengan menggunakan algoritma klasterisasi dapat menunjukkan bahwa algoritma ini berhasil memiliki fungsi linear terhadap waktu dan pengurangan ruang.

Dardac dan Boitan (2009) juga menyatakan bahwa pengelompokan dengan menggunakan klasterisasi dapat membantu permasalahan untuk melakukan analisa kredit terhadap suatu lembaga keuangan. Dengan adanya metode ini, lembaga keuangan dapat mengurangi resiko sistematis yang ada, diantaranya kebangkrutan kreditur. Akan tetapi berbeda dengan penelitian sebelumnya (Oyelade, Oladipupo, dan Obagbuwa, 2010; Dragut, 2012; Andayani, 2007; dan Dardac dan Boitan, 2009), Pham, Dimov, dan Nguyen (2005) menyebutkan bahwa meskipun Algoritma K-Means merupakan algoritma yang populer untuk klasterisasi, jumlah klaster menentukan kinerja dari Algoritma ini. Hal ini didukung oleh peneliti – peneliti lainnya (Rakhlin dan Caponnetto, 2013) yang menyebutkan bahwa keakuratan hasil klasterisasi dipengaruhi oleh overlapping yang terjadi pada metode K-Means.

Metode K-means juga memberikan beban komputasi yang tinggi (Murugesan dan Zhang, 2011). Murugesan dan Zhang (2011) menyarankan suatu metode baru untuk mengurangi beban komputasi, yaitu dengan menggabungkan kombinasi *divisive* dan *agglomerative hierarchical klasterisasi (K-Means dan Unweighted Pair Group Method with Arithmetic Mean (UPGMA)).* Sebagai hasilnya, dengan menggunakan algoritma hybrid beban komputasi dapat di kurangi.

Jarak eucledian juga menjadi penentu dalam menentukan kestabilan sebuah klaster. Jika semakin kecil jarak Eucledian suatu klaster, maka kesamaan yang dihasilkan semakin bagus dan membuat klaster yang dihasilkan lebih proporsional (Sriparna dan Bandyopadhayay, 2009). Berbeda dengan Sriparna dan Bandyopadhayay (2009), Chiang dan Mirkin (2010) menyatakan bahwa penentuan jumlah klaster menjadi permasalahan yang tersendiri dalam metode ini. Hal ini juga didukung oleh Tapas, et. al (2002). Oleh karena itu, Chiang dan Mirkin (2010) menggunakan metode iK-Means yang merupakan gabungan metode Hartigan untuk menemukan jumlah klaster yang cocok dengan jumlah kluster K dalam K-Means. Sebagai kesimpulannya, penulis menyebutkan bahwa metode Hartigan dapat digunakan untuk menentukan jumah k dalam metode K-Means. Berbeda dengan Chiang dan Mirkin (2010), Tapas, et. al. (2002) menggunakan metode Lloyd untuk menentukan jumlah klaster k dalam

sekumpulan data R^d. Perhitungan jumlah klaster membutuhkan suatu perhitungan kd-tree yang digunakan untuk merepresentasikan subdivisi hierarki dari suatu klaster.

Untuk mendapatkan kualitas kinerja yang baik dari kerangka kerja klasterisasi, Bilgina, et. al. (2010) menggunakan tahap preprocessing untuk meningkatkan performa. Adapun pendekatan ini digunakan untuk melihat efektivitas dan efisiensi dari masing – masing algoritma yang ada. Didapatkan suatu metode untuk dataset yang lebih besar sehingga dapat meningkatkan hubungan antardataset yang ada dan meningkatkan skalabilitas dari dimensi yang dimiliki.

Chen, Ching, dan Lin (2004) menyebutkan bahwa algoritma K-Means sebagai salah satu algoritma klasterisasi yang paling banyak digunakan, terutama untuk penelitian di bidang pemasaran. Chen, et. al (2004) mencoba untuk menerapkan metode K-Means dengan menerapkan konsep hierarchical klasterisasi untuk meningkatkan kualitas dari solusi yang dihasilkan. Sebagai hasilnya, algoritma yang dihasilkan memiliki kemampuan lebih baik jika dibandingkan dengan metode K-Means lainnya.

Ding dan He (2004) menggunakan metode K-Means dan *Principal Component Analysis* untuk melakukan pengelompokan terhadap newsgroup yang terdapat di Internet. *Principal Component Analysis* merupakan salah satu metode yang efektif untuk mengurangi dimensi dari suatu klaster. Penggunaan dua metode ini secara bersamaan juga dilakukan oleh Alexe, et. al. (2007) dalam melakukan analisa terhadap kanker payudara. Alexe, et. al (2007) menggunakan metode *Principal Component Analysis* untuk melakukan identifikasi keseluruhan dari suatu data dan memilih klaster dari subset yang ada. Setelah didapatkan pembawa sifat keturunan dari penyakit tersebut, metode K-Means diterapkan untuk melakukan identifikasi terhadap klaster yang kuat dan stabil.

Selain metode K-Means, beberapa penulis mencoba untuk memberikan suatu teknik klasterisasi dengan menggunakan metode kesamaan dalam tekstur pola yang ada dari tiap data (Kheradmandian dan Rahmati, 2009). Kheramandian dan Rahmati (2009) menggunakan konsep texture berdasarkan kepadatan atau

kekurangan dari suatu pola yang menjadi milik suatu klaster tertentu. Texture tersebut digunakan untuk mencari pola yang terdapat dalam suatu klaster. Metode ini lebih bagus dibandingan dengan algoritma *agglomerative hierarchical klasterisasi*.

Berbeda dengan Kheradmandian dan Rahmati (2009), Strehl dan Ghosh (2003) menyatakan bahwa metode klaster untuk data dengan dimensi tinggi dapat menggunakan metode *Relationship Based Klasterisasi*. Di dalam *Relationship Based Klasterisasi*, penggunaan kesamaan perantara dapat digunakan untuk merepresentasikan klaster dari tiap – tiap pelanggan yang ada. Dengan menggunakan visualisasi akan sangat membantu dalam melakukan penilaian dan memperbaiki kinerja dari proses *klasterisasi*. Penyusunan kembali data points yang ada dapat menggunakan kesamaan dalam *similarity matrix* (Strehl dan Ghosh, 2003).

Penggunaan metode K-Means untuk melakukan pengelompokan seringkali dipadukan dengan metode lainnya untuk mendapatkan pengetahuan yang ada dari tiap klaster, diantaranya dengan Jaringan Syaraf Tiruan. Penggunaan Jaringan Syaraf Tiruan untuk melakukan penilaian terhadap resiko adalah salah satu penerapan Jaringan Syaraf Tiruan di bidang ekonomi (Ramamoorti, et. al., 1999; Vojinovic, Kecman, dan Seidel, 2001). Ramamoorti, et. al. (1999) menggunakan Backpropagation Neural Network untuk melakukan penilaian terhadap suatu resiko. Penulis membandingkan dengan metode statistika yang lainnya. Penulis memaparkan bahwa penelitian yang digunakan merupakan metode yang efektif untuk melakukan penilaian terhadap suatu resiko (Ramamoorti, et. al., 1999). Vojinovic, Kecman, dan Seidel (2001) menggunakan pendekatan Radial Basis Function Neural Network untuk melakukan prediksi nilai tukar Dollar Selandia Baru terhadap dollar Amerika (US$). Vojinovic, Kecman, dan Seidel (2001) menyatakan bahwa *Radial Basis Function Neural Network* dapat memberikan keakuratan terhadap hasil prediksi. Hal ini disebabkan Jaringan Syaraf Tiruan dapat menemukan jumlah jaringan yang tersembunyi. Selain itu, metode Jaringan syaraf tiruan juga dapat di kombinasikan dengan algoritma genetik untuk melakukan analisis efektivitas frontier (Azadeh, et. al., 2009).

Metode Jaringan Syaraf Tiruan merupakan metode yang seringkali digunakan untuk melakukan prediksi, diantaranya untuk melakukan prediksi kegagalan suatu usaha (Shah dan Murtaza, 2000), prediksi terhadap volume paru – paru (Manoharan dan Ramakrishnan, 2009), dan peramalan di bidang keuangan (Pavlidis, et. al., 2006). Pada penelitiannya, Shah dan Murtaza (2000) menyebutkan bahwa gabungan metode klasterisasi dengan jaringan syaraf tiruan memberikan tingkat prediksi yang sangat tinggi, yaitu 73%. Sehingga gabungan kedua metode ini sangat efektif untuk melakukan prediksi (Shah dan Murtaza, 2000; Manoharan dan Ramakrishnan, 2009). Penggabungan metode klasterisasi dengan Jaringan Syaraf tiruan memberikan hasil yang bagus untuk melakukan prediksi terhadap daerah yang berbeda (Pavlidis, et. al., 2006).

Penggunaan pengabungan *Principal Component Analysis* dengan Jaringan Syaraf juga dilakukan oleh Zee (2011) untuk melakukan perhitungan menjadi lebih efisien. Beberapa metode yang digunakan yaitu : (1) Klasifikasi dengan menggunakan minor Principal Componenet; (2) Rotasi dari principal component sebelum dilakukan dengan jaringan syaraf tiruan; (3) turunan dari dua atau lebih *principal component analysis* dan jaringan syaraf tiruan untuk klasterisasi lithofacies. Sebagai hasilnya *principal component analysis* memiliki kemampuan yang efektif untuk menyeleksi komponen – komponen menjadi lebih sedikit (Zee, 2011).

Romdhane, Fadhel, dan Ayeb (2009) menyimpulkan bahwa teknologi klasterisasi dengan menggabungkan Fuzzy Klasterisasi dengan *Backpropagation Neural Network* mampu memberikan simulasi yang bagus untuk melakukan pengelompokan terhadap data pelanggan. Selain itu, dengan adanya *Backpropagation Neural Network* mampu memberikan suatu kecerdasan kepada sistem sehingga dapat memberikan pengetahuan dari kumpulan data pelanggan yang ada.

Berbeda dengan Romdhane, Fadhel, dan Ayeb (2009), penggunaan klasterisasi penambangan data juga digunakan di dalam dunia medis. Chen, et. al (2011) menggunakan pendekatan hybrid penambangan data dengan menggabungkan Klasterisasi dan pohon keputusan (decision tree) untuk

melakukan pencegahan terhadap kesalahan dalam pemberian obat (*drug dispensing error*). Sebagai hasilnya, Chen, et. al. (2011) menemukan bahwa Hybrid Penambangan data yang digunakan efektif untuk memberikan peringatan kepada pengguna jika ada kesalahan pada saat pemberian obat.

Gao, et. al (2005) menggunakan pemberian rekomendasi dalam perpustakaan digital dengan menggunakan klasterisasi dan *partial back propagation*. Penulis menggunakan *frequent pattern* yang ada di dalam suatu grup untuk mendapatkan pola – pola yang sering ditemui. Pola yang ada digunakan sebagai data pelatihan dengan menggunakan metode *Partial Back Propagation Neural Network*. Gao, et. al. Menyimpulkan bahwa gabungan metode klasterisasi dengan *Partial Back Propagation Neural Network* memberikan dampak yang sangat efektif untuk memberikan rekomendasi kepada pelanggan.

Berdasarkan paparan tinjauan pustaka sebelumnya, penulis mencoba untuk menggunakan metode *K-Means* klasterisasi dan *Principal Component Analysis* untuk menggelompokkan data saham. Setelah klaster terbentuk, data ini akan digunakan sebagai data pelatihan bagi Jaringan Syaraf Tiruan dengan menggunakan metode *Back Propagation Neural Network*.

B. Landasan Teori

1 *Simple Moving Average* (SMA)

Simple Moving Average adalah salah satu metode unweighted untuk menghitung mean dari n buah data. Di dalam penulisan ini, penulis menggunakan metode *Simple Moving Average* untuk melengkapi data yang hilang. Data yang hilang diambilkan dari 5 data sebelum dan 5 data sesudahnya. *Simple Moving Average* dapat di hitung sebagai:

$$SMA = \frac{X_i + X_{i+1} + X_{i+2} + ... + X_{(i+n-1)}}{n} \qquad \text{--- (2.1)}$$

Dimana:

SMA : *Simple Moving Average* dari n buah data

N : Banyaknya data

X_i : Data suku ke i

2 Penambangan data

Penambangan data adalah suatu proses iterasi yang ditentukan dengan suatu penemuan, baik itu secara otomatis ataupun secara manual (Mehmed, 2003). Dengan menggunakan penambangan data, suatu informasi yang baru, bernilai, dan informasi – informasi lain yang dapat di tentukan dari sekumpulan data yang besar. Pada dasarnya, terdapat dua prinsip dasar yang ada pada penambangan data, yaitu :

a) Prediksi

Pada prediksi, penambangan data digunakan untuk membuat suatu model yang dapat digunakan untuk melakukan prediksi berdasarkan data set yang dimiliki.

b) Deskripsi

Pada deskripsi, penambangan data digunakan untuk menemukan suatu informasi yang baru berdasarkan data set yang dimiliki.

Menurut Mehmed(2003), terdapat beberapa teknik penambangan data yang sering digunakan, yaitu :

a) Klasifikasi

Teknik ini digunakan untuk menemukan suatu fungsi prediksi yang mengelompokkan data menjadi satu atau lebih kelompok.

b) Regresi

Metode regresi digunakan untuk menemukan suatu fungsi prediksi dengan memetakan data menjadi sebuah fungsi yang terdiri dari beberapa variabel, yaitu *dependent* dan *independent* variabel - variabel. Nilai dari suatu dependent variabel selalu tergantung dari satu atau lebih independent variabel.

c) Klasterisasi

Klasterisasi adalah metode yang digunakan untuk mengelompokkan data set menjadi beberapa kelompok (klaster) berdasarkan prinsip kesamaan antaranggota dalam satu klaster.

d) Peringkasan

Metode ini digunakan untuk menemukan suatu deskripsi dari satu set data. Dengan metode ini, fungsi agregrat akan diterapakan untuk menemukan suatu deskripsi dari data set yang tersedia.

e) Model Ketergantungan

Metode ini digunakan untuk menemukan ketergantungan yang signifikan antarvariabel atau antardata yang terdapat dalam suatu dataset.

f) Deteksi Perubahan dan Deviasi

Metode ini digunakan untuk menemukan perubahan – perubahan yang signifikan terhadap suatu dataset.

3 Analisa Klaster

Teknik *klasterisasi* merupakan salah satu teknik yang cukup di kenal dan banyak digunakan dalam ilmu *penambangan data*. Banyak penelitian di kembangkan dan di laksanakan untuk mendapatkan cara yang efektif dalam melakukan klasterisasi.

Sebuah klaster adalah sekumpulan data yang memiliki kesamaan (*similar*) dengan data yang lain di dalam klaster yang sama dan tidak memiliki kesamaan (*dissimilar*) kepada objek – objek yang lain di dalam klaster yang berbeda. Dengan menggunakan metode klasterisasi, objek – objek yang sama dapat di kelompokkan menjadi satu untuk memudahkan identifikasi.

Tujuan utama dari metode klasterisasi adalah pengelompokan sujumlah data / obyek ke dalam suatu group sehingga dalam setiap klaster akan berisi data yang semirip mungkin. Dalam klasterisasi, obyek – obyek yang mirip di di tempatkan dalam satu klaster dan membuat jarak antarklaster sejauh mungkin. Ini berarti obyek dalam satu klaster sangat mirip satu sama lain dan berbeda dengan obyek dalam klaster – klaster yang lain. Selain itu, dengan menggunakan klasterisasi, beberapa faktor yang mempengaruhi perbedaan antara klaster satu dengan lainnya dapat ditentukan. Metode klasterisasi telah diimplementasikan untuk beberapa bidang, diantaranya kesehatan, geografi, bisnis, dan beberapa bidang lainnya.

Di dalam dunia bisnis, klaster analysis digunakan untuk penelitian pangsa pasar, penentuan *supplier*. Selain itu, klasterisasi juga dapat membantu tim marketing untuk menemukan karakteristik untuk tiap konsumen dari tiap daerah berdasarkan kebiasaan membeli. Hal ini akan membantu dalam mengetahui produk apa yang dibutuhkan oleh konsumen dan bagaimana menyusun strategi pemasaran yang efektif untuk tiap daerah.

Sedangkan di dunia kesehatan, beberapa ahli menggunakan klasterisasi untuk mengklasifikasikan daerah rawan penyakit. Dengan demikian, akan membantu dinas kesehatan dalam penanganan suatu wabah penyakit dan menentukan bagimana strategi yang efektif untuk menangani suatu penyebaran penyakit.

Klasterisasi disebut juga dengan *data segmentation* di beberapa applikasi. Dengan klasterisasi, sekumpulan dataset yang sangat besar, dikelompokkan / disegmentasikan menjadi beberapa group, berdasarkan kesamaan yang dimiliki oleh tiap - tiap anggota group. Selain itu, klasterisasi juga membantu untuk mendeteksi outlier (sekumpulan data yang sangat berbeda jauh dari kelompok manapun).

Terdapat dua pendekatan yang biasa di lakukan dalam klasterisasi, yaitu partisi dan hirarki. Dalam pastisi ini, kita mengelompokkan obyek x_1, x_2, x_3, x_4, ..., x_n ke dalam *k* klaster. Hal ini bisa di lakukan dengan menentukan pusat klaster awal, lalu di lakukan realokasi obyek berdasarkan kriteria tertentu sampai di capai pengelompokkan yang optimum. Dalam klaster hirarki di mulai dengan membuat *m* klaster dimana setiap klaster beranggotakan satu obyek dan berakhir dengan satu klaster dimana anggotanya adalah *m* obyek. Pada setiap tahap dalam prosedurnya, satu klaster di gabungkan dengan satu klaster yang lain.

Pada klasterisasi hirarki, jarak masing – masing obyek di hitung dengan setiap obyek yang lain sehingga akan ditemukan pasangan obyek yang jaraknya terdekat. Sehingga tiap obyek akan berpasangan dengan satu obyek atau kelompok obyek yang lain yang paling dekat jaraknya. Dengan demikian, metode klasterisasi hirarki menghitung jarak kemiripan (*similarity*) antarobyek.

4 *Similarity* dan *Disimilarity*

Di dalam klasterisasi, untuk menggabungkan dua atau lebih obyek menjadi satu klaster, biasanya digunakan ukuran kemiripan atau ketidakmiripan. Semakin mirip dua obyek, semakin tinggi peluang untuk dikelompokkan menjadi satu klaster. Sebaliknya semakin tidak mirip, semakin rendah peluang untuk di kelompokkan dalam satu klaster.

Untuk ukuran kemiripan dapat digunakan cosinus, kovarian, dan korelasi. Dalam ukuran kemiripan, jika dua obyek atau lebih memiliki nilai yang semakin tinggi, maka dua obyek atau lebih tersebut semakin mirip. Terdapat tiga (3) cara yang dapat di gunakan untuk menentukan kemiripan antara obyek, yaitu :

a) Cosinus

Cosinus antara dua titik x dan y di definisikan sebagai berikut:

$$\cos^{-} = \frac{x^T y}{\|x\|\|y\|} \qquad \text{--- (2.2)}$$

Dimana:

$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^{\,2}} \qquad \text{---(2.3)}$$

b) Kovarian

Kovarian antara dua data (x dan y) di definisikan sebagai berikut:

$$Cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \qquad \text{---(2.4)}$$

c) Korelasi

Sedangkan untuk korelasi antara dua data (x dan y) dapat di definisikan sebagai berikut:

$$r(x, y) = \frac{Cov(x, y)}{\pm_x \pm_y} \qquad \text{---(2.5)}$$

Di dalam melakukan perhitungan similarity terdapat beberapa macam tipe data yang disebut dengan *multivariate* tipe data, seperti nominal, ordinal, dan kuantitatif harus diolah terlebih dahulu menjadi data numerik. Pengubahan data

numerik ini ditujukan agar algoritma K-Means dapat digunakan dalam menggunakan algoritma K-Means.

Atribut yang berbeda tipe berarti terdapat ketidaksamaan (*dissimilarity*) antaratribut tersebut. Ketidaksamaan antara dua obyek dapat diukur dengan menghitung jarak antarobyek berdasarkan beberapa sifatnya. Ketidaksamaan dapat di lihat sebagai berikut :

a) $d(a,b) \geq 0$,jarak kedua obyek selalu positif atau nol
b) $d(a,a) = 0$, jarak terhadap diri sendiri adalah nol
c) $d(a,b) = d(b,a)$, jarak kedua obyek adalah simetri
d) $d(a,b) \leq d(a,c) + d(c,b)$, jarak memenuhi ketidaksamaan segitiga.

Hubungan antara kesamaan dan ketidaksamaan juga dapat di asumsikan sebagai berikut

$$S_{i,j} = 1 - d_{i,j} \qquad \text{--- (2.6)}$$

Dimana :

$S_{i,j}$: Kesamaan antara obyek i dan obyek j

$d_{i,j}$: Ketidaksamaan antara obyek i dan obyek j

5 K-Means Klasterisasi

Algoritma K-Means paling banyak digunakan dalam berbagai penelitian dan dalam banyak industri. Algoritma ini dimulai dengan menetapkan nilai acuan yang akan digunakan sebagai titik tengah dari klaster (*Cluster Centroid Point*). Algoritma untuk melakukan metode K-Means Klasterisasi adalah sebagai berikut:

a) Menentukan Jumlah klaster yang akan di buat.
b) Inisialisasi nilai pusat klaster (k)
c) Tempatkan setiap obyek ke dalam klaster terdekat. Kedekatan dua obyek di tentukan dengan melakukan perhitungan kesamaan antara obyek tersebut
d) Hitung kembali pusat klaster dengan anggota klaster sekarang. Pusat klaster adalah rata – rata semua data / obyek dalam satu klaster.

e) Kelompokkan kembali pusat klaster dengan anggota klaster yang baru. Pusat klaster adalah rata – rata semua data / obyek dalam satu klaster.

f) Ulangi langkah 3 sampai pusat klaster tidak berubah.

Berikut ini adalah diagram alir (*flowchart*) untuk metode K-Means Klasterisasi :

Gambar 2.1: Diagram Alir Metode K-Means

6 Principal Component Analysis

Pengurangan dimensi (*dimension reduction*) yang terdapat pada sekelompok data digunakan untuk mengurangi beberapa variable yang tidak berelasi antarsatu variabel dengan variabel lainnya dalam satu kelompok klaster. Tujuan dari

pengurangan dimensi adalah untuk mendapatkan variabel – variabel yang optimal yang dapat membentuk klaster yang diinginkan. Salah satu metode pengurangan dimensi adalah *Principal Component Analysis*.

Principal Component Analysis menurut Lindsay (2002) adalah sebuah metode untuk mengidentifikasi pola – pola yang terdapat dalam sebuah data dan menyatakannya dalam sebuah cara untuk menentukan kemiripan dan perbedaan yang dimiliki oleh data tersebut. Salah satu keunggulan yang dapat ditemukan dalam *Principal Component Analysis* adalah dengan melakukan metode ini dapat mengurangi jumlah dimensi yang terdapat dalam satu pola tanpa mengurangi informasi yang terdapat dalam data tersebut. Oleh karena itu, *Principal Component Analysis* sangat diperlukan untuk membantu reduksi terhadap pola – pola yang ada dalam suatu klaster.

Principal Component Analysis sangat cocok untuk digunakan terhadap high-dimensional dataset. PCA lebih dekat dikenal dengan analisa faktor (factor analysis). Beberapa langkah yang digunakan untuk menggunakan metode *Principal Component Analysis* adalah :

a) Persiapkan data yang akan dianalisa dengan menggunakan *Principal Component Analysis*.
b) Hitung *mean* untuk kelompok data tersebut.
c) Melakukan perhitungan untuk matrik kovarian, dimana varian dihitung sebagai :

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{(n-1)} \qquad \text{--- (2.7)}$$

Dimana :

s^2 : kovarian data

N : jumlah data

X_i : data ke-i

$\overline{X}$: rata – rata semua data

Dan kovarian dihitung sebagai :

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)} \quad \text{---- (2.8)}$$

Dimana :

X_i : koordinat X data ke-i

$\overline{X}$: rata – rata X semua data

Y_i : koordinat Y data ke-i

$\overline{Y}$: rata – rata Y semua data

Kovarian matrik dihitung sebagai:

Jika kita memiliki data dengan dimensi (x, y, z), kita dapat melakukan kalkulasi $Cov(X,Y)$, $Cov(X,Z)$, $Cov(Y,Z)$. Dengan kata lain, apabila kita memiliki data dengan dimensi n, maka banyaknya kovarian matrik yang kita hitung yaitu : $\frac{n!}{(n-2)!*2}$

d) Melakukan perhitungan eigenvectors dan eigenvalues dari matrik kovarian.

e) Mendapatkan dataset yang baru berdasarkan variabel – variabel yang terkait dalam satu faktor.

7 Jaringan Syaraf Tiruan

Jaringan Syaraf Tiruan adalah suatu metode yang dapat memberikan suatu metode yang disusun sesuai dengan kinerja jaringan syaraf yang terdapat dalam susunan struktur tubuh manusia (Budi, 2007). Terdapat beberapa karakteristik kemampuan otak manusia yang mendasari metode Jaringan Syaraf Tiruan, yaitu : mengingat, menghitung, mengeneralisasi, dan adaptasi. Oleh karena itu metode ini diharapkan sebagai alternatif pendekatan konvensional yang biasanya kurang fleksibel terhadap perubahan struktur masalah. Jaringan Syaraf Tiruan menawarkan kelebihannya dimana dapat mengatasi beberapa persoalan tanpa mengadakan perubahan drastis terhadap model. Salah satu metode yang ada dalam jaringan syaraf tiruan adalah algoritma back-propagasi.

Algoritma back-propagasi sangat bermanfaat,cukup handal, dan mudah dipahami. Error adalah selisih antara target yang sebenarnya dengan keluaran dari jaringan yang ada. Error ini dapat di ketahui dengan menggunakan rumusan sebagai berikut :

$$e_j(n) = d_j(n) - y_j(n) \quad \text{--- (2.9)}$$

Sehingga jumah keseluruhan error untuk semua unit keluaran dapat dinotasikan sebagai berikut :

$$E(n) = \frac{1}{2}\sum_j e_j^2(n) \quad \text{--- (2.10)}$$

Sedangkan menurut gradient descent learning rule, w di update dengan menggunakan formula sebagai berikut :

$$w_{ji}(n+1) = w_j(n) + \Delta w_{ji}(n) \quad \text{--- (2.11)}$$

Sedangkan $\Delta w_{ji}(n)$ adalah :

$$\Delta w_{ji}(n) = -\eta \frac{\partial E(n)}{\partial w_{ji}} \quad \text{--- (2.12)}$$

Dimana E adalah error training dengan menggunakan perhitungan :

$$E(w) = \frac{1}{2}\sum_{k=1}^{c}(d_k - y_k)^2 \quad \text{--- (2.13)}$$

Dimana :

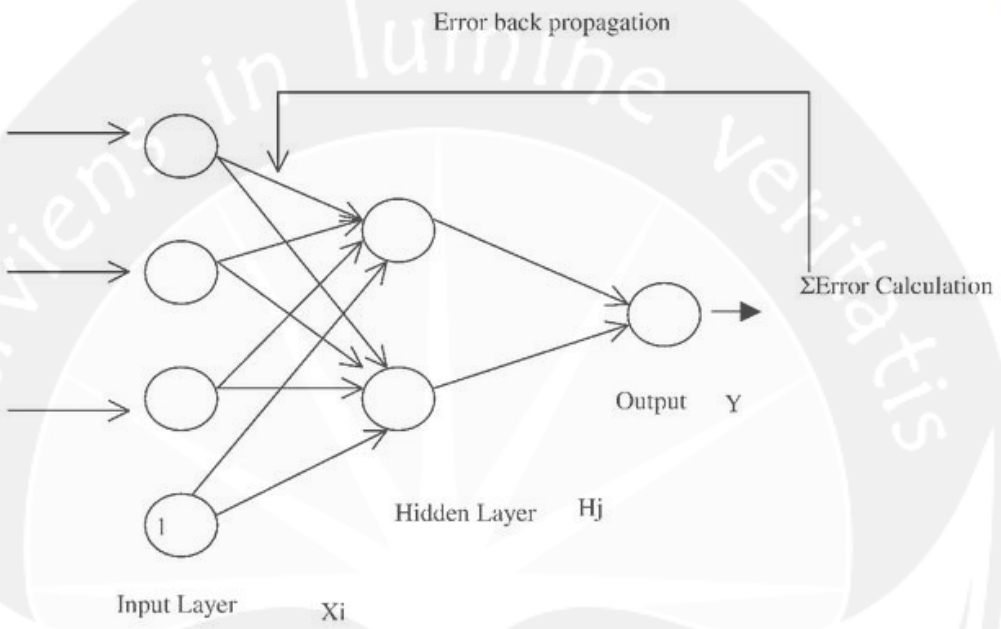
d_k : nilai target untuk unit k

y_k : keluaran sebenarnya untuk unit k

w_{ji} : Bobot yang berhubungan dengan input ke – i ke unit j.

Arsitektur Backpropagation

Backpropagation memiliki beberapa unit yang ada dalam satu atau lebih layer tersembunyi. Arsitektur backpropagation digambarkan pada gambar 2.2 berikut ini dengan n buah masukan (X_i), sebuah layer tersembunyi (*Hidden Layer* H_j) yang terdiri dari 2 unit, dan 1 buah unit keluaran (Output Y).



Gambar 2.2 : Arsitektur Backpropagation[4]

Pada gambar 2.2 diatas, bobot V_{ji} dapat dikatakan sebagai bobot garis dari unit masukan X_i ke unit layer tersembunyi H_j. Sedangkan bobot W_{yj} merupakan bobot garis dari unit layer tersembunyi H_j kepada keluaran Y yang terdapat pada jaringan backpropagation.

[4] Data diambil dr http://www.emeraldinsight.com/journals.htm?articleid=1614232&show=html. Data diakses pada tanggal 23 September 2013.