BAB II

LANDASAN TEORI

2.1 Data Mining

Secara sederhana, data mining merupakan ekstraksi informasi yang tersirat dalam sekumpulan data. Data mining merupakan sebuah proses untuk menggali kumpulan data dan menemukan informasi di dalamnya. (Turban, dkk, 2005).Data mining merupakan proses pengekstrakan informasi dari jumlah kumpulan data yang besar dengan menggunakan algoritma dan tehnik gambar dari statistik, mesin pembelajaran dan system manajemen database. Penggalian data ini dilakukan pada sekumpulan data yang besar untuk menemukan pola atau hubungan yang ada dalam kumpulan data tersebut (Kusrini & Luthfi, 2009). Hasil penemuan yang diperoleh setelah proses penggalian data ini, kemudian dapat digunakan untuk analisis yang lebih lanjut.

Data mining yang disebut juga dengan *Knowledge-Discovery in Database* (KDD) adalah sebuah proses secara otomatis atas pencarian data di dalam sebuah memori yang amat besar dari data untuk mengetahui pola dengan menggunakan alat seperti klasifikasi, hubungan (*association*) atau pengelompokan (*clustering*). Proses KDD ini terdiri dari langkah-langkah sebagai berikut (Han, J. & Kamber, M, 2001):

1. *Data Cleaning*, proses menghapus data yang tidak konsisten dan kotor
2. *Data Integration*, penggabungan beberapa sumber data
3. *Data Selection*, pengambilan data yang akan dipakai dari sumber data
4. *Data Transformation*, proses dimana data ditransformasikan menjadi bentuk yang sesuai untuk diproses dalam data mining
5. *Data Mining*, suatu proses yang penting dengan melibatkan metode untuk menghasilkan suatu pola data

6. *Pattern Evaluation*, proses untuk menguji kebenaran dari pola data yang mewakili knowledge yang ada didalam data itu sendiri
7. *Knowledge Presentation*, proses visualisasi dan teknik menyajikan knowledge digunakan untuk menampilkan knowledge hasil mining kepada user.

2.2 Informasi Data Set

Data set merupakan data yang kita ambil dari *data warehouse* yang nantinya akan kita proses ulang dengan metode-metode tertentu untuk diteliti lebih lanjut supaya nanti didapatkan informasi yang baru. Sebuah data set biasanya berbentuk seperti tabel, di mana baris-barisnya menyatakan observasi (pengamatan/individu) dan kolom-kolomnya merupakan variabel. Struktur data set adalah sama dengan data base relasional, karena menghadapkan model objek hirarkis tabel, baris, dan kolom. Selain itu, berisi kendala dan hubungan didefinisikan untuk data set.

2.2.1 *Clustering* data

Clustering atau analisis *cluster* adalah proses pengelompokan satu set benda-benda fisik atau abstrak ke dalam kelas objek yang sama (Han and Kamber, 2006).

Clustering atau *clusterisasi* adalah salah satu alat bantu pada data *mining* yang bertujuan mengelompokkan obyek-obyek ke dalam *cluster-cluster*. *Cluster* adalah sekelompok atau sekumpulan obyek-obyek data yang *similar* satu sama lain dalam *cluster* yang sama dan *dissimilar* terhadap obyek-obyek yang berbeda *cluster*. Obyek akan dikelompokkan ke dalam satu atau lebih *cluster* sehingga obyek-obyek yang berada dalam satu *cluster* akan mempunyai kesamaan yang tinggi antara satu dengan lainnya.

Dengan menggunakan *clusterisasi*, kita dapat mengidentifikasi potensi siswa.

2.3 Pengertian Metode SOM

Self Organizing Map (SOM) merupakan metode pengelompokan dalam bentuk topografi dua dimensi layaknya sebuah peta sehingga memudahkan pengamatan distribusi hasil pengelompokan. SOM memerlukan penentuan laju pembelajaran, fungsi pembelajaaran, jumlah iterasi yang di inginkan dalam proses pengelompokannya. untuk memberikan hasil pengelompokan Metode *Self Organizing Map* tidak memerlukan fungsi objektif seperti *K-Means* dan *Fuzzy C-Means* sehingga untuk kondisi yang sudah optimal pada suatu iterasi, SOM tidak akan menghentikan iterasinya selama jumlah iterasi yang ditentukan belum tercapai.

SOM *(Self Organizing Feature Map)* pertama kali diperkenalkan oleh Kohonen (Kohonen,1989) dengan teknik pelatihan ANN yang menggunakan *basis winner takes*

all. Di mana hanya neuron yang menjadi pemenang yang akan diperarui bobotnya. Meskipun menggunakan basis ANN, SOM tidak menggunakan target kelas, tidak ada kelas yang ditetapkan untuk setiap data. Karakteristik seperti inilah yang kemudian membuat SOM dapat digunakan untuk keperluan pengelompokan berbasis ANN (Prasetyo, 2012).

2.4 Algoritma SOM

Berikut ini adalah langkah-langkah yang perlu dilakukan dalam menerapkan metode SOM dalam pengolahan data (Prasetyo, 2012):

1. Inisialisasi Nilai bobot Wij secara acak, tentukan parameter topologi ketetanggaan, tentukan parameter laju pembelajaran, tentukan jumlah iterasi pelatihan.

2. Selama jumlah maksimal iterasi belum tercapai, lakukan langkah 3 -7.

3. Untuk setiap data masukan X (matriks M x N), lakukan langkah 4 – 6.

4. Untuk setiap neuron j, hitung $D_j = \sum_i \left(w_{ij} - x_i\right)^2$, i=1,...,N,N adalah dimensi data (N).

5. Cari indeks dari sejumlah neuron, yaitu D_j, yang mempunyai nilai terkecil.

6. Untuk neuron j dan semua neuron yang menjadi tetangga J (yang sudah didefinisikan) dalam radius R, hitunglah perubahan bobot Wij(baru) = Wij(lama) + ή (X_i – W_{ij}(lama)).

7. Perbaharui laju pembelajaran dengan rumus :
   $\eta(baru) = fungsi\ pembelajaran \times \eta(lama)$

Pada algoritma tersebut dapat dijelaskan, parameter jarak untuk perbedaan atau kemiripan yang digunakan adalah Encludian kuadrat (square Euclidean). Hal ini dimaksudkan untuk mengurangi waktu komputasi dengan menyederhanakan kinerja algoritma, tetapi harus dibayar dengan penggunaan memori yang lebih besar untuk alokasi nilai jarak yang biasnya lebih besar. Nilai laju pembelajaran (..) yang digunakan adalah jangkauan nilai 0 sampai 1. Tetapi, nilai ini akan terus diturunkan setiap kali ada kenaikan iterasi dengan sebuah fungsi pembelajaran.

Inisialisasi bobot awal bisa menggunakan nilai acak dengan jangkauan -0.5 sampai +0.5, atau menggunakan nilai acak dengan jangkauan nilai seperti pada data masukan.

Parameter yang bisa digunakan dalam pengelompokan set data X pada sejumlah C. untuk penjelasanya dapa dilihat pada tabel 2.1 :

Tabel 2.1 : Parameter dan keterangan algoritma SOM.

No	Parameter	Keterangan
1.	X	X adalah matriks MxN. M menyatakan jumlah data, dan N menyatakan jumlah fitur.
2.	C	C menyatakan jumlah kelompok (neuron pemroses)
3.	Iterasi	Jumlah maksimal iterasi pelatihan

4.	W_{ij}	Matriks NxC yang menyatakan bobot untuk setiap neuron (kelompok). N menyatakan jumlah fitur, dan C menyatakan jumlah neuron (kelompok).
5.	D_j	Matriks MxC. Baris menyatakan data, dan kolom menyatakan jarak ke neuron (kelompok).
6.	J	J menyatakan neuron keluaran yang didapat dari data yang berdimensi tinggi. Neuron keluaran tersebut berfungsi jika dalam penelitian menggunakan topologi SOM.

2.5 Penelitian sebelumnya

1. Penelitian sebelunya dilakukan berorientasi obyek yaitu berdasarkan data-data yang sudah tersimpan seperti nilai-nilai bidang studi dan nilai tes IQ sebelumnya kemudian digunakan sebagai acuan data dan perhitungan dengan menggunakan metode SOM yang hasilnya nanti digunakan sebagai acuan penjurusan siswa yang berdasarkan hasil nilai studi.

   "Diagnosa potensi peserta didik", Oleh Hamid Muhammad, Ph.D. kasus permasalahan pada paper ini yaitu, pembelajaran yang dilaksanakan selama ini bersifat massal, yang memberikan perlakuan dan layanan pendidikan yang sama kepada semua peserta didik. Hasil dari kesimpulan paper ini yaitu setiap siswa mempunyai tingkat kecakapan, kecerdasan, minat, bakat, dan kreativitasnya yang berbeda-beda. Strategi pelayanan pendidikan seperti ini memang tepat dalam konteks pemerataan kesempatan, akan tetapi kurang menunjang usaha mengoptimalkan pengembangan potensi peserta didik secara cepat [HAM04]. Kelebihan dan kekurangan dari paper ini yaitu, Strategi pelayanan pendidikan alternatif dalam manajemen pendidikan perlu dikembangkan untuk menghasilkan peserta didik yang

unggul, melalui pemberian perhatian, perlakuan dan layanan pendidikan berdasarkan bakat minat dan kemampuannya. Agar pelayanan pendidikan yang selama ini diberikan kepada peserta didik mencapai sasaran yang optimal, maka pembelajaran harus diselaraskan dengan potensi peserta didik. Oleh karena itu guru perlu melakukan pelacakan potensi peserta didik

2. Penelitian yang dilakukan salah seorang dosen pada Universitas Negeri Padang dengan judul penelitian "**Evaluasi Materi Bahan Ajar Berbasis Multimedia pada Mata Kuliah Listening 1 di Jurusan Bahasa dan Sastra Inggris**". Materi bahan ajar ini diterapkan pada tahun 2011 merupakan materi bahan ajar yang lama dan perlu diadakan perbaikan bahan ajar tersebut. Untuk itu diperlukan adanya evaluasi guna mengetahui apakah materi yang baru sesuai untuk mahasiswa, dosen dan institusi tersebut. Evaluasi materi ajar ini mencakup pendekatan dan metodologi, desain dan organisasi, tata bahasa, sub-keahlian menyimak, materi multimedia, topik dan latihan.

   Metode yang digunakan pada penelitian ini adalah metode penelitian dengan sistem evaluasi yang diterapkan pada mahasiswa, dari hasil penelitian yang dilakukan pada universitas tersebut menunjukkan bahwa materi ajar berbasis multimedia pada mata kuliah Listening 1 sesuai dan cocok untuk mahasiswa, dosen dan institusi. Selain itu, diperoleh pula informasi mengenai kekurangan dari materi yang dianggap perlu untuk ditingkatkan. Dari hasil penelitian tersebut metode yang diterapkan bisa diterima oleh mahasiswa, dosen dan intitusi tersebut.