BAB II

TINJAUAN PUSTAKA DAN LANDASAN TEORI

Pada bab ini akan dibahas literatur dan landasan teori yang relevan dengan penelitian.

2.1 Tinjauan Pustaka

Kombinasi metode telah dilakukan oleh beberapa peneliti sebelumnya dengan algoritma yang berbeda-beda seperti yang telah dipaparkan pada tabel 1.1. Berikut penjelasan singkat algoritma yang membedakan setiap metode kombinasi pada setiap penelitian:

Xie dkk (2002) menggabungkan metode KNN dengan Naïve Bayes. Metode utama yang digunakan ialah metode Naïve Bayes dimana KNN digunakan pada langkah pertama untuk mencari jarak antara data uji dengan setiap data pelatihan. Kemudian dilanjutkan dengan proses Naïve Bayes menggunakan jarak antara data uji dengan setiap data pelatihan dan dilanjutkan dengan mencari nilai probabilitas terbesar sesuai dengan tahapan pada Naïve Bayes. Tujuan dari munculnya algoritma baru ini untuk meningkatkan nilai keakuratan pada penambangan data, dan terbukti dapat meningkatkan nilai keakuratan dari Naïve Bayes.

Jiang dkk (2005) menggabungkan metode KNN dengan Naïve bayes. Metode utama yang digunakan ialah metode KNN dimana Naïve Bayes digunakan pada saat tahap perhitungan peringkat jarak terdekat metode KNN dan

mencloned beberapa data pelatihan yang memiliki nilai jarak terdekat dengan data uji. Kemudian mencari nilai probabilitas dari data tersebut dengan Naïve Bayes. Tujuan dari munculnya algoritma baru ini untuk meningkatkan nilai keakuratan pada penambangan data, dan terbukti bahwa hasil yang diperoleh lebih baik dibandingkan dengan Naïve Bayes, C4.4 dan NBTree.

Hall (2007), menggabungkan Naïve Bayes dengan Decision Tree. Metode utama yang digunakan ialah Naïve Bayes dimana Decision Tree digunakan pada awal pembobotan atribut. Tujuan munculnya algoritma baru ini untuk meningkatkan proses kinerja Naïve Bayes berdasarkan algoritma Decision Tree yang terbukti lebih baik dibandingkan dengan Naïve Bayes trees dan selective Bayes.

Farid dkk (2010), menggabungkan metode Naïve Bayes dengan Decision Tree. Metode utama yang digunakan ialah Naïve Bayes, dan Decision Tree digunakan untuk mencari Gain data pada tahap klasifikasi. Langkah pertama ialah mencari nilai probabilitas dengan menggunakan Naïve Bayes dan nilai tersebut digunakan untuk memperbarui nilai setiap kelas pada D. Kemudian memilih gain terbaik dengan nilai maksimum. Tujuan dari munculnya algoritma baru ini untuk mendeteksi gangguan jaringan dengan nilai akurat yang tinggi. Hasil dari percobaan yang diujikan, keakuratan algoritma ini dalam mendeteksi gangguan jaringan mencapai 99%.

2.2 Landasan Teori

2.2.1. Data, Informasi dan Pengetahuan

Data adalah bilangan, terkait dengan angka- angka atau atribut-atribut yang bersifat kuantitas yang berasal dari hasil observasi, eksperimen, atau kalkulasi. Data kategori adalah semua nilai yang mungkin ada, bersifat terbatas yang berdasarkan nominal dan ordinal. Yang dimaksud nominal di sini adalah tanpa adanya urutan sebagai contohnya adalah status perkawinan atau jenis kelamin. Informasi adalah data didalam satu konteks tertentu. Informasi merupakan kumpulan data dan terkait dengan penjelasan, interpretasi, dan berhubungan dengan materi lainnya mengenai objek, peristiwa-peristiwa atau proses tertentu. Sementara itu, pengetahuan adalah informasi yang telah diorganisasi, disintesiskan, diringkaskan untuk meningkatkan pengertian, kesadaran atau pemahaman (Bergeron, 2003).

2.2.2. Klasifikasi Penambangan data

Penambangan data yang juga dikenal dengan KDD (*Knowledge Discovery in Database*) digunakan untuk mengekstrak model yang menggambarkan data kelas yang penting (Baradwaj & Pal, 2011, Bhargavi & Jyothi, 2011, Bhuvaneswari & Kalaiselvi, 2012, Kumar & Verma, 2012, Kaur & Aggarwal, 2013, Kaur & Kaur, 2013).

Sebelum melakukan proses penambangan data, perlu dilakukan beberapa tahapan yang disebut *preprocessing* (Beniwal & Arora, 2012). Hal tersebut disebabkan karena teknik preprocessing memiliki dampak signifikan terhadap

kinerja pada algoritma machine learning. Desain database yang baik dan analisis yang baik dapat mereduksi permasalahan missing data melalui preprocessing (Christobel & Sivaprakasam, 2013). Dengan menggunakan teknik visualisasi preprocessing dapat diperoleh beberapa pengetahuan mengenai data (Nithyasri dkk, 2010). Berikut merupakan beberapa tahapan awal hingga hasil dari penambangan data (Singh dkk, 2012):

1. Pemilihan (*data selection*)

   Pemilihan data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai.

2. Pemrosesan awal (*preprocessing*)

   Sebelum proses penambangan data dapat dilaksanakan, perlu dilakukan proses cleaning dengan tujuan untuk membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Juga dilakukan proses enrichment, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. Transformasi

   Proses coding pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses penambangan data. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam database.

4. Penambangan data (*data mining*)

   Proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam penambangan data sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. Interpretasi/Evaluasi

   Pola informasi yang dihasilkan dari proses penambangan data perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut dengan interpretation.Tahap ini mencangkup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya atau tidak.

2.2.3. Teknik Klasifikasi

Klasifikasi merupakan tugas penambangan data yang memetakan data ke dalam kelompok-kelompok kelas (Jain dkk, 2013). Teknik klasifikasi melakukan pengklasifikasian item data ke label kelas yang telah ditetapkan, membangun model klasifikasi dari kumpulan data input, membangun model yang digunakan untuk memprediksi tren data masa depan (Shazmeen dkk, 2013). Algoritma yang umum digunakan meliputi *K-Nearest neighbor*, *Naïve Bayes Classification*, Pohon Keputusan (*Decision Tree*), Jaringan Saraf (*Neural Network*), dan *Suport Vector Machines* (Sahu dkk, 2011).

2.2.2.1 Naïve Bayes

Naïve Bayes merupakan metode klasifikasi yang statistik berdasarkan teorema Bayes (Kabir dkk, 2011, Baby & T., 2012). Naïve Bayes berpotensi baik untuk mengklasifikasikan data karena kesederhanaannya (Abraham dkk, 2009, Ting dkk, 2011).

Persamaan yang digunakan pada Naïve Bayes:

$$P\ c_i\ x\ = \frac{P\ x\ c_i\ P\ c_i}{P_x} \quad (2.1)$$

P(c_i|X) yaitu, probabilitas c_i terjadi jika X sudah terjadi.

P(c_i) adalah kemungkinan c_i didata, bersifat independent terhadap X.

X adalah kumpulan atribut.

P(X| c_i) adalah probabilitas X terjadi jika c_i benar atau sudah terjadi berdasarkan data pelatihan.

Selama nilai P(x) konstan, maka dapat disederhanakan menjadi rumus berikut ini:

$$P\ C_i\ X\ = P\ X\ C_i\ P(C_i) \quad (2.2)$$

2.2.2.2 KNN (K-Nearest Neighbor)

KNN (K-Nearest Neighbor) merupakan metode yang cukup popular dan sederhana (Chou &Shen, 2006). KNN termasuk metode klasifikasi penambangan data yang didasarkan pada pembelajaran dengan analogi. Sampel data pelatihan memiliki n atribut dimensi numerik. Setiap sampel merupakan titik dalam ruang n-dimensi. Semua sampel pelatihan disimpan di ruang n-dimensi. Ketika pengujian data, akan mencari nilai k terdekat dengan data uji. Kedekatan

didefinisikan dalam hal jarak Euclidean antara dua titik X=(x_1, x_2,…, x_n) dan Y=(y_1, y_2,…, y_n) (Phyu, 2009).

$$d\ X,Y\ = \ \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (2.3)$$