KONSTRUKSI BAYESIAN NETWORK DENGAN ALGORITMA K2 PADA KASUS PREDIKSI CUACA

Herastia Maharani [#1]

[#] *Departemen Teknik Informatika*
Institut Teknologi Harapan Bangsa
Telp: +62 250 6636
[1] herastia@ithb.ac.id

Abstrak— Sistem peramalan cuaca dibangun dengan menganalisis hubungan antar variabel cuaca. Salah satu pendekatan untuk menganalisis hubungan antar variabel cuaca adalah dengan menggunakan data mining. Bayesian Network merupakan salah satu metode data mining yang dapat menggambarkan hubungan sebab-akibat antara variabel dalam sebuah sistem. Dalam penelitian ini dibangun Bayesian Network dengan algoritma K2 untuk memodelkan hubungan antara variabel-variabel cuaca dan melakukan prediksi berdasarkan model yang dihasilkan. Hasil pengujian menunjukkan bahwa Bayesian Network mampu memodelkan hubungan antar variabel cuaca dan menghasilkan prediksi yang cukup akurat.

Kata kunci— variabel cuaca, Bayesian Network, algoritma K2, prediksi

Abstract— Weather forecast systems are built by analyzing the dependency between weather variables. One alternative approach to analyze the dependency between weather variables is by using data mining technique. Bayesian Network is a method capable of representing causal dependencies between variables in a system. This research used Bayesian Network built by using K2 algorithm to model the dependencies between weather variables and making predictions based on the obtained model. The experiment results showed that Bayesian Network managed to represent the dependencies between weather variables and made quite accurate predictions.

Keywords— weather variables, Bayesian Network, K2 algorithm, predictions

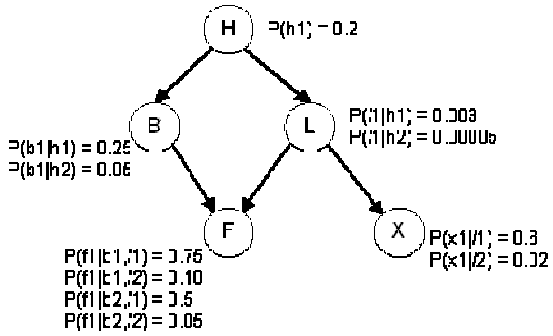
I. PENGANTAR

Cuaca merupakan hal yang tidak pernah bisa lepas dari kehidupan manusia. Kondisi cuaca dapat mempengaruhi jalannya aktivitas manusia, sebagai contoh tingkat curah hujan dapat mempengaruhi keputusan manusia untuk pergi beraktivitas di luar rumah, keputusan petani dalam menggarap lahan pertanian, persiapan menghadapi banjir, dan sebagainya. Besarnya pengaruh faktor cuaca ini mendorong berkembangnya sistem peramalan cuaca yang mencoba memprediksi kondisi cuaca di masa depan. Sistem peramalan cuaca yang ada saat ini menerapkan berbagai pendekatan dengan menggunakan dukungan teknologi yang ada.

Salah satu pendekatan yang digunakan dalam memprediksi kondisi cuaca adalah dengan memanfaatkan konsep data mining. Data mining digunakan untuk menganalisis dan menemukan pola yang terdapat dalam perubahan cuaca dan memodelkan keterhubungan antar variabel cuaca. Dari model yang dihasilkan, dapat dilakukan perhitungan untuk memprediksi kondisi atau nilai dari variabel cuaca yang ingin dketahui, misalnya tingkat curah hujan. Salah satu metode data mining yang umum digunakan untuk memodelkan keterhubungan antar variabel adalah *Bayesian Network*. Tujuan dari penelitian ini adalah untuk membangun model yang merepresentasikan hubungan antar variabel cuaca dengan menggunakan Bayesian Network. Model yang dihasilkan akan digunakan untuk memprediksi nilai variabel tertentu (missal curah hujan) berdasarkan kondisi variabel yang lain. Algoritma yang digunakan untuk membangun Bayesian Network dalam penelitian ini adalah algoritma K2 yang dijelaskan dalam [1, 2].

II. BAYESIAN NETWORK (BN)

Salah satu cara merepresentasikan *pattern* dalam *data mining* adalah dengan menggunakan model grafis yang dinamakan probabilistic graphical model (PGM). Bayesian Network (BN) merupakan salah satu bentuk PGM yang digunakan untuk merepresentasikan pattern dalam data mining. Pearl menyatakan bahwa BN adalah graf asiklik berarah (directed acyclic graph, disingkat DAG) dimana setiap node merepresentasikan variabel dan setiap arc merepresentasikan pengaruh sebab-akibat di antara variabel [3]. Definisi lain dari Cheng menyatakan bahwa BN adalah sebuah DAG dengan tabel probabilitas untuk setiap node di dalamnya [4]. Setiap node dalam BN merepresentasikan variabel proposisional dalam sebuah domain, dan arc yang ada antar node merepresentasikan hubungan kebergantungan antar variabel tersebut [4]. Dalam mengkonstruksi BN dari basis data, maka node digunakan untuk merepresentasikan atribut-atribut dalam basis data. Contoh Bayesian Network diberikan pada gambar 1.



Gambar 1: Contoh Bayesian Network [5]

Pada contoh di gambar 1, setiap *node* merepresentasikan sebuah variabel atau atribut dalam basis data. *Edge* yang menghubungkan dua buah *node* merepresentasikan adanya hubungan sebab-akibat antara kedua *node* tersebut. Dapat dilihat bahwa pada setiap *node* terdapat nilai probabilitas yang menunjukkan peluang dari setiap kemungkinan nilai untuk *node* tersebut, relatif terhadap nilai *parent*-nya. Struktur DAG yang terdiri dari *node* dan *edge* merupakan komponen struktur dari Bayesian Network, sedangkan nilai probabilitas di setiap *node* merupakan komponen parameter dari bayesian Network.

Secara umum terdapat dua pendekatan dalam mempelajari struktur BN dari data, yaitu metode *dependency analysis* dan metode *search and scoring*. Penelitian ini menggunakan salah satu algoritma *search and scoring* yaitu algoritma K2.

III. ALGORITMA K2

Untuk menyelesaikan kasus ini, akan digunakan algoritma K2 yang dikembangkan oleh Cooper dan Herzkowits [1, 2]. Cara kerja algoritma K2 diberikan pada Gambar 2 [1, 2]:

```
1.  for each node i, 1 ≤ i ≤ n, find πi
    as follows
2.      πi ← Φ
3.      pOld ← f(i, πi )
4.      notDone ← false

5.      while not notDone do
6.          Let z be sets of pred(xi)
7.          pNew ← f(i, πi ∪ (z)
8.          if pNew > pOld then
9.              pOld = pNew
10.             πi ← πi ∪ (z)
11.         else
12.             notDone = true
13.         end if
14.     end while
```

Gambar 2: Algoritma K2

Algoritma K2 merupakan algoritma yang menggunakan metode *search and* scoring, artinya algoritma ini menggunakan sebuah fungsi *scoring* untuk mengevaluasi struktur *network* yang dibangkitkan. Fungsi *scoring* yang digunakan dalam algoritma ini adalah *Bayesian Scoring Function* yang dinyatakan dalam persamaan Cooper-Herskovits berikut [1, 2]:

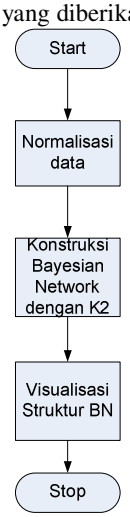
$$f(i,\pi_i) = \prod_{j=1}^{n} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)} \prod_{k=1}^{ri} \alpha_{ijk}!$$

IV. METODOLOGI PENELITIAN

Dalam penelitian ini digunakan enam buah variabel cuaca, yaitu suhu (*temperature*), kelembaban relatif (*relative humidity*), tingkat penguapan (*evaporation*), kuantitas awan (*cloud amount*), tekanan permukaan laut (*MeanSeaLevel Pressure*), dan curah hujan (*Rain*). Data yang digunakan merupakan data nilai keenam variabel cuaca tersebut selama kurang lebih satu tahun yang diperoleh dari [6].

Untuk memodelkan keterhubungan antara keenam variabel ini, dibangun sebuah aplikasi untuk membangun struktur Bayesian Network dari data yang diperoleh dengan menggunakan algoritma K2. Model yang dihasilkan akan digunakan untuk memprediksi nilai salah satu variabel, yaitu curah hujan, berdasarkan nilai variabel-variabel yang lain.

Secara umum alur kerja dari sistem yang dibangun dapat dilihat pada Gambar 3. Terdapat tiga tahapan, yaitu normalisasi data, konstruksi Bayesian Network, dan visualisasi model Bayesian Network. Tahap normalisasi bertujuan untuk membagi data yang bersifat kontinyu ke dalam kelas-kelas yang representati. Tahap konstruksi Bayesian Network berisi implementasi dari algoritma K2 dan menghasilkan struktur DAG dan parameter Bayesian Newtork. Sedangkan tahap terakhir adalah memvisualisasikan DAG yang dihasilkan ke pengguna. Fokus utama dalam penelitian ini adalah tahap kedua, yaitu implementasi algoritma K2 untuk menghasilkan struktur dan parameter Bayesian Network yang mampu merepresentasikan hubungan antar variabel cuaca yang digunakan. Implementasi algoritma K2 dilakukan sesuai dengan algoritma yang diberikan di Gambar 2.



Gambar 3: Tahapan Utama dalam Sistem

V. HASIL PENGUJIAN

Pengujian dalam penelitian ini dibagi menjadi dua tahap sebagai berikut:

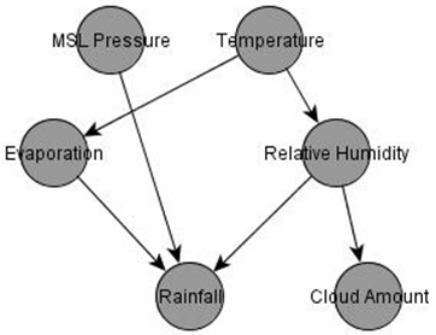
(i) Mengevaluasi struktur Bayesian Network yang dihasilkan dengan algoritma K2.

(ii) Mengevaluasi parameter Bayesian Network yang dihasilkan. Pengujian dilakukan dengan menggunakan model yang dihasilkan untuk memprediksi nilai curah hujan. Hasil prediksi kemudian dibandingkan dengan nilai aktualnya untuk megevaluasi akurasi prediksi yang dihasilkan dari model yang dibangun.

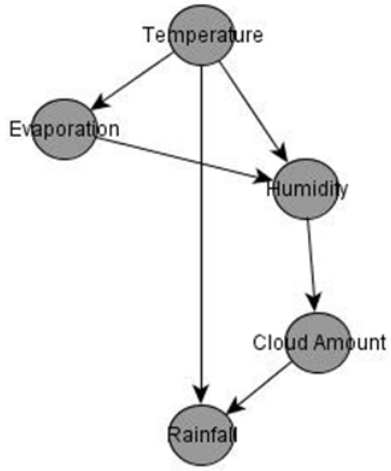
A. *Evaluasi Struktur Bayesian Network*

Proses konstruksi Bayesian Network dilakukan dengan menggunakan data cuaca untuk keenam variabel yang sudah disebutkan sebelumnya. Data yang digunakan terbagi dalam dua kelompok, masing-masing menggambarkan nilai untuk dua kota yang berbeda. Hasil struktur yang dihasilkan dapat dilihat pada Gambar 4 dan Gambar 5.

Dapat dilihat bahwa untuk kota A (Gambar 4), keenam variabel yang digunakan muncul dalam DAG yang dihasilkan dan dapat dilihat bahwa memang terdapat keterhubungan antar keenam variabel tersebut. Contoh hubungan yang ditemukan adalah bahwa nilai variabel *Relative Humidity* dipengaruhi secara langsung oleh nilai dari variabel *Temperature*. Sedangkan nilai variabel *Rainfall* dipengaruhi secara langsung oleh variabel *Evaporation*, *Relative Humidity*, dan *MSLPressure*. Artinya, kita dapat memprediksi nilai *Relative Humidity* jika kita mengetahui nilai dari *Temperature*. Adapun nilai variabel *Rainfall* dapat diprediksi berdasarkan nilai variabel *Evaporation*, *MSLPressure*, dan *Relative Humidity*.



Gambar 4: Struktur Bayesian Network dari Data Kota A



Gambar 5 : Struktur Bayesian Network dari Data Kota B

Adapun untuk struktur Bayesian Network yang dihasilkan dari data kota B, hanya ada lima variabel yang muncul dalam DAG yang dihasilkan (Gambar 5). Variabel *MSLPressure* tidak muncul dalam struktur DAG yang dihasilkan. Hal ini kemungkinan besar disebabkan karena dalam data kota B, nilai dari variabel *MSLPressure* tidak memiliki hubungan yang signifikan terhadap nilai variabel yang lain. Tidak ada variabel lain yang nilainya berpengaruh pada nilai *MSLPressure* secara signifikan, sehingga pada saat dilakukan proses *scoring* tidak ditemukan *parent* untuk *MSLPressure*. Di lain pihak, nilai dari variabel *MSLPressure* juga tidak member pengaruh terhadap nilai variabel lain, sehingga tidak ada variabel lain yang menjadi *child* dari *MSLPressure*. Hubungan antar variabel yang dihasilkan untuk kita B pun memiliki perbedaan dengan kota A. Untuk kota B, nilai *Rainfall* dipengaruhi oleh nilai *CloudAmount* dan *Temperature*.

Perbedaan struktur yang dihasilkan dari kedua kota ini disebabkan oleh perbedaan kondisi nilai keenam variabel cuaca di kedua kota tersebut. Seperti halnya teknik *data mining* pada umumnya, model yang dihasilkan dengan Bayesian Network sangat bergantung pada dataset yang digunakan dalam proses pembelajaran dan konstruksi model. Perbedaan kondisi dan distribusi nilai variabel cuaca di kedua kota tersebut mengakibatkan struktur DAG yang dihasilkan untuk kedua kota tersebut memiliki perbedaan.

B. *Evaluasi Parameter Bayesian Network*

Untuk mengevaluasi nilai parameter yang dihasilkan, dilakukan pengujian dengan memprediksi nilai *Rainfall* berdasarkan nilai variabel *parent*-nya. Data yang digunakan sebagai data uji adalah data yang diambil dari [6] yang tidak digunakan sebagai dataset dalam proses konstruksi Bayesian Network. Hasil prediksi yang dihasilkan akan dibandingkan dengan nilai aktual curah hujan yang tercatat dalam data uji. Karena nilai actual curah hujan berupa nilai kontinyu, maka hasil prediksi hanya dibagi menjadi dua kelas saja, yaitu Tidak Hujan (*no rain*) dan Hujan (*rain*). Prediksi yang dihasilkan belum mampu menghitung berapa tepatnya curah hujan yang

terjadi. Beberapa contoh hasil prediksi untuk data kota A dapat dilihat pada Tabel 1.

Kolom Curah Hujan Aktual menunjukkan data aktual curah hujan yang diambil dari data uji, sedangkan kolom Prediksi menunjukkan persentase probabilitas setiap kelasnya. Kolom Kelas 1 dan Kelas 2 menunjukkan probabilitas terjadinya setiap kelas yang dihitung berdasarkan struktur dan parameter Bayesian Network yang dihasilkan untuk kota A (Gambar 4). Kolom Status Prediksi memberikan keterangan apakah hasil prediksi yang diberikan dianggap valid atau tidak. Dari contoh hasil pengujian yang diberikan di Tabel 1 dapat dilihat bahwa semua hasil prediksi berstatus *Valid*. Hal ini dikarenakan sebagian besar hasil prediksi mengarah ke kelas yang sesuai dengan nilai aktual curah hujan untuk kasus uji tersebut. Contohnya untuk kasus uji 3 dan 4, hasil prediksi memilih Kelas 1 (*no rain*), prediksi ini sesuai dengan curah hujan aktualnya yang bernilai 0 (nol), artinya untuk kasus itu memang tidak ada hujan. Namun demikian, terdapat beberapa kasus uji yang menunjukkan perbedaan antara hasil prediksi dengan data aktual, contohnya pada kasus uji 1 dapat dilihat bahwa prediksi menunjukkan tidak hujan, namun fakta mengatakan hujan. Hasil ini tetap dinyatakan *Valid* karena setelah dilakukan analisis terhadap data uji tersebut, ternyata data ini terjadi pada musim kemarau dimana kemunculan hujan memang sangat jarang. Artinya, tetap ada kemungkinan bahwa tidak semua hari menunjukkan tidak hujan, pasti ada minimal 1 hari yang hujan. Fenomena seperti ini dalam data mining disebut dengan *outlier*. Karena itu, prediksi untuk kasus ini masih bersifat *Valid* mengingat tingkat level hujan yang tidak begitu besar. Dari total 62 kasus uji yang digunakan, model yang dihasilkan dapat memberikan prediksi dengan akurasi sebesar 75.80%.

TABEL 1: HASIL PENGUJIAN PREDIKSI *RAINFALL*

Kasus Uji	**Curah hujan aktual**	**Hasil prediksi (%)**		**Status Prediksi**
		Kelas 1 (*no rain*)	**Kelas 2 (*rain*)**	
1	1.2	66.7	33.3	Valid
2	5.6	0	100	Valid
3	0	100	0	Valid
4	0	71.5	28.5	Valid
5	0	40	60	Valid
6	3.8	100	0	Valid
7	0	100	0	Valid
8	0	83.3	16.7	Valid
9	0	80	20	Valid
10	0	88.9	11.1	Valid
11	29	0	100	Valid
12	0	100	0	Valid
13	0	88.8	11.1	Valid

Dari kedua pengujian di atas dapat disimpulkan bahwa *Bayesian Network* dapat memodelkan hubungan sebab akibat antar variabel cuaca dengan cukup baik. Hasil prediksi yang dihasilkan pun memiliki akurasi yang cukup tinggi. Hanya saja, model yang dihasilkan memang sangat bergantung pada dataset yang digunakan dalam proses konstruksi Bayesian Network. Perbedaan dalam datayang digunakan akan sangat berpengaruh pada struktur yang dihasilkan. Semakin lengkap dan representatif data yang digunakan, maka model Bayesian Network dan prediksi yang dihasilkan akan semakin akurat.

VI. KESIMPULAN

1. Penggunaan algoritma K2 terbukti mampu menghasilkan Bayesian Network yang cukup reprsentatif dengan akurasi prediksi yang cukup baik.
2. Kuantitas dataset dan distribusi nilai variabel dalam dataset yang digunakan dalam proses konstruksi Bayesian Network memiliki pengaruh yang signifikan terhadap struktur dan akurasi prediksi yang dihasilkan.
3. Struktur Bayesian Network yang dihasilkan sangat dipengaruhi oleh kondisi data yang digunakan. Struktur yang dihasilkan dari satu dataset belum tentu berlaku pada dataset lainnya, Dalam kasus cuaca ini, struktur yang dihasilkan untuk satu kota belum tentu berlaku untuk kota lainnya.

REFERENSI

[1] G. F. Cooper and E. A. Herskovits, "A Bayesian Method for the Induction of Probabilistic Networks from Data", Machine Learning 9, 309-347, 1992.

[2] C. Ruiz, "Illustrations of the K2 Algorithm for Learning Bayes Net Structures", Lecture Notes on Machine Learning, Department of Computer Science, Worcester Polytechnic Institute, 2005.

[3] J. Pearl, "Graphical Models for Probabilistic and Causal Reasoning". Computer Science Department, University of California, 1997.

[4] J. Cheng, D. Bell, W. Liu, "Learning Bayesian Networks from Data: An Efficient Approach Based On Information Theory", Faculty of Informatics, University of Ulster, U.K., 1998

[5] R. E. Neapolitan, "Learning Bayesian Networks", Pearson Prentice Hall, 2004

[6] http://www.bom.gov.au, diakses tanggal 9 Desember 2009