IMPLEMENTASI METODE *CENTROID DECOMPOSITION* UNTUK PERAMALAN PADA DATA CUACA

Irfan Pratama
Fakultas Teknologi Informasi, Program Studi Sistem Infromasi
Universitas Mercubuana Yogyakarta
Irfanp@mercubuana-yogya.ac.id

***Abstrak*— Data mining adalah sebuah fase pencarian pengetahuan pada kumpulan suatu data. Data mining juga adalah sebuah proses ekstraksi dari informasi-informasi dan pengetahuan-pengetahuan yang berguna yang didapat dari kumpulan data yang besar, tidak lengkap, acak, dan ambigu. Berdasarkan pengetahuan tersebut, penelitian ini dilakukan untuk mengetahui apakah metode yang diterapkan oleh peneliti sebelumnya pada penanganan missing values dapat diterapkan pada proses prediksi dengan beberapa penyesuaian. Seiring bertambahnya titik prediksi, hasil dari metode Ekstrapolasi Linear semakin buruk. Dengan kata lain tidak cocok untuk melakukan prediksi jangka menengah hingga panjang, namun dapat dilakukan menggunakan metode Centroid Decomposition.**

***Kata Kunci*— Centroid Decomposition, Ekstrapolasi Linear, Data Cuaca, Prediksi.**

I. PENDAHULUAN

Pada era digital seperti saat ini, penggalian informasi pengetahuan dan informasi melalui data-data menjadi suatu yang sangat penting(Rahman & Islam 2016). Data mining adalah sebuah fase pencarian pengetahuan pada kumpulan suatu data. Data mining juga adalah sebuah proses ekstraksi dari informasi-informasi dan pengetahuan-pengetahuan yang berguna yang didapat dari kumpulan data yang besar, tidak lengkap, acak, dan ambigu(Han et al. 2011). Data mining didefinisikan sebagai bentuk analisis data secara otomatis atau semi-otomatis dari dataset yang besar dan kompleks untuk mendapatkan informasi berupa pola, kaitan atau hubungan antar data (Han et al. 2011).

Salah satu pemanfaatan proses data data mining adalah pada proses prediksi cuaca. Cuaca sendiri adalah suatu hal yang bersifat alami dan tidak dapat secara pasti diprediksi.meskipun demikian, banyak studi-studi dan penelitian atau bahkan penggunaan prediksi cuaca yang digunakan dalam kehidupan sehari-hari. Salah satu contohnya adalah prakiraan cuaca harian yang disiarkan atau dikabarkan oleh media-media cetak maupun digital. Proses data mining terhadap cuaca juga dapat berupa prediksi ketersediaan air yang berasal dari seberapa seberapa besar curah hujan pada suatu area atau kondisi area tersebut kedepannya dengan mempelajari data-data historis dari elemen-elemen cuaca seperti, suhu udara, intensitas curah hujan, kelembaban relatif, dan lain-lain(Omary et al. 2012).

Contoh lain dari pemanfaatan prediksi terhadap data cuaca adalah pada bidang agrikultur. Setiap pengetahuan terhadap pola dan hasil prediksi data cuaca akan berguna untuk estimasi jenis tanaman yang baik ditanam hingga estimasi profit yang akan didapat dari proses cocok tanam tersebut(Shivaranjani 2016). Karakteristik data cuaca yang cenderung berpola membuat proses pendekatan data mining atau prediksi menjadi cukup terbantu. Akan tetapi sifatnya yang alami membuat kondisi-kondisi diluar dugaan kerap muncul seperti kejadian anomali pada cuaca yang disebabkan oleh perubahan iklim. Sehingga membuat proses kajian data-data historis dari data cuaca tersebut menjadi tidak mudah. Secara teknis, kondisi-kondisi cuaca dapat direkam oleh alat-alat observasi yang kemudian direprsentasikan menjadi angka-angka.

Pada konteks matematis, terdapat dua pendekatan yang dapat digunakan untuk menghasilkan nilai-nilai baru dari sekumpulan nilai, yaitu Interpolasi dan ekstrapolasi. Interpolasi merupakan pemunculan nilai-nilai diantara nilai yang sudah ada, dan ekstrapolasi adalah metode pemunculan nilai setelah nilai-nilai yang ada. Pada sebuah penelitian yang dilakukan oleh (Borzsonyi 2013), digunakan sebuah metode dasar yang berbasis interpolasi yang digunakan untuk menangani masalah missing values. jika dilihat dari konsep penanganannya, missing values dan prediksi merupakan dua hal yang mirip, yaitu sama-sama memprediksi nilai menggunakan nilai-nilai yang ada. Perbedaanya hanya terletak pada posisi dimana nilai-nilai prediksi tersebut ditempatkan.

Berdasarkan pengetahuan tersebut, penelitian ini dilakukan untuk mengetahui apakah metode yang diterapkan oleh (Borzsonyi 2013) pada penanganan missing values dapat diterapkan pada proses prediksi dengan beberapa penyesuaian.

II. METODOLOGI

Penelitian ini dilakukan dengan membuat beberapa skenario simulasi untuk mengukur sejauh mana hasil dari metode yang digunakan pada penelitian ini.

Pada bagian ini akan dijelaskan tentang alur penelitian, metode-metode yang digunakan, dan model evaluasi yang akan digunakan dan penjelasan masing-masing bagiannya adalah sebagai berikut.

A. *Olah Data Awal*

Penelitian ini menggunakan data suhu udara harian daerah Melbourne dari rentang tahun 1986-1990, data pada penelitian ini diperoleh secara terbuka dari www.datamarket.com. Data yang diperoleh dalam kondisi sempurna dan lengkap. Sampel dari data yang akan digunakan pada penelitian ini dapat dilihat pada Tabel I. Untuk melakukan simulasi pada penelitian ini

mula-mula akan dilakukan beberapa skenario penghapusan data untuk kemudian dilakukan prediksi terhadap data sejumlah data yang dihilangkan tersebut. Skenario-skenario penghilangan data yang dilakukan pada penelitian ini dapat dilihat pada Tabel II.

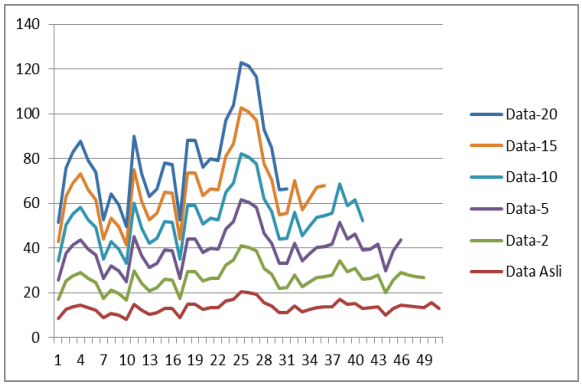
TABEL I
SAMPEL DATA PENELITIAN

Hari Ke-	Data Suhu	Hari Ke-	Data Suhu	Hari Ke-	Data Suhu
1	14.8	6	15.3	11	22.1
2	13.3	7	16.4	12	19
3	15.6	8	14.8	13	15.5
4	14.5	9	17.4	14	15.8
5	14.3	10	18.8	15	14.7

TABEL III
SKENARIO PENGHILANGAN DATA

Skenario	Data yang di prediksi
1	2
2	5
3	10

Pada Tabel I dapat dilihat bahwa data suhu udara yang digunakan dalam bentuk bilangan desimal. Sedangkan pada Tabel II menunjukan berapa banyak data yang akan di prediksi. Sebagai ilustrasi data awal dan data hasil rekayasa dapat dilihat pada Gambar 1.



Gbr. 1 Ilustrasi Data

Pada Gambar 1 dapat dilihat perbedaan panjang data dari data awal yang berjumlah lengkap yang kemudian dihilangkan secara berkala sesuai dengan skenario pada Tabel II dimana data dihilangkan berurutan dari mulai data yang paling akhir hingga jumlah yang ditentukan.

B. *Metode Prediksi*

Metode utama yang digunakan pada penelitian ini sebagian besar adalah metode yang disebut Centroid Decomposition. Metode Centroid Decomposition adalah metode berbasis dekomposisi matriks yang proses nya bersifat iteratif hingga mencapai kondisi tertentu (Borzsonyi 2013). Pada setiap iterasinya, metode ini menghasilkan Centroid Factor dan LoadingVector. Berikut adalah prosedur dalam bentuk Pseudocode pada Gambar 2.

Algoritme : *Centroid Decomposition* (A)
Input : A^{nxm}
Output : *loading matrix B* dan *factor matrix V*
1. i = 1;
2. $A_i := A$;
3. ***Repeat***
4. z = *FindSignVector*(A_i)
5. $c_i = A_i^T z$;
6. $v_i = \frac{c_i}{\|c_i\|}$;
7. $b_i = A_i v_i$;
8. $A_i = A_i - b_i v_i^T$;
9. If i = 0 then
10. $B = b_i, V = v_i$;
11. Else
12. B = append(B,b_i) //meletakkan b_i pada sisi kanan vektor B
13. V = append(V,v_i) ////meletakkan v_i pada sisi kanan vektor V
14. i = i +1;
15. m = m – 1;
16. ***Until*** m = 0;
End of Pseudo-Code

Gbr. 2 Metode *Centroid Decomposition*

Gambar 2 menunjukan prosedur utama dari metode Centroid Decomposition. Pada prosedur tersebut terdapat fungsi FindSignVector() yang merupakan bagian lebih detail dari metode tersebut. Prosedur untuk fungsi FindSignVector() dapat dilihat pada Gambar 3.

Algoritme : *FindSignVector* (A)
Input : A
Output : *sign vector* z untuk matriks A
1. *pos* = 0;
2. **Repeat**
3. // merubah tanda
4. If *pos* = 0 then $z^T = [1, \cdots, 1]$;
5. else rubah tanda pada z_{pos};
6. endif
7. $S = \sum_{i=1}^{n} (z_i \times (X_{pi})^T)$;
8. $V = []$;
9. For i = 1 to *n* do
10. $v_i = X_{pi}.S - z_i \times X_{pi}.(X_{pi})^T$;
11. Masukkan v_i pada *V*;
12. endfor
13. // mencari elemen selanjutnya
14. *val* = 0, *pos* = 0;
15. for *i* = 1 to *n* do
16. If ($z_i * v_i < 0$) then
17. If $|v_i| >$ *val* then
18. *val* = v_i;
19. *pos* = *i*;
20. endif
21. endif
22. endfor
23. **Until** *pos* = 0;
24. **Return** z;

Gbr. 3 Fungsi *FindSignVector*

Pada Gambar 3 diperlihatkan proses yang berjalan pada fungsi FindSignVector guna mendapatkan komponen yang akan digunakan untuk prosedur utama pada Gambar 2.

Secara sistematis, prosedur bagaimana sebuah data yang akan di prediksi mulai di proses hingga mendapat hasil akhir prediksi adalah sebagai berikut:

1) Prediksi Awal: Setiap data yang yang akan menjalani proses prediksi menggunakan metode Centroid Decomposition akan terlebih dahulu di prediksi menggunakan model prediksi yang sederhana (contoh: ekstrapolasi linear) unutk mendapatkan nilai prediksi awal yang kemudian akan diolah didalam metode utamanya.

2) Pemilihan data referensi: Setelah sebuah baris data di prediksi menggunakan metode extrapolasi linear akan dipotong pada interval dimana datanya akan menjadi dua kali dari panjang data yang di prediksi (contoh: data yang di prediksi = 2 maka, panjang vektor data hasil potongan = 4 data) dengan komposisi separuh dari baris data tersebut adalah data-data yang berdekatan dengan data hasil prediksi. Selanjutnya, baris data dengan panjang tertentu tersebut akan diukur korelasinya dengan data-data aktual historis pada tahun-tahun data sebelumnya. Setelah didapatkan nilai korelasi menggunakan perhitungan korealsi Pearson, diambilah vektor dengan korelasi tertinggi dan terendah (data dengan korelasi tertinggi berfungsi sebagai nilai acu utama sedangkan baris data dengan korelasi terendah berfungsi sebagai acuan bentuk data). Kemudian ketiga vektor tersebut dibentuk menjadi sebuah matriks dengan dimensi nx3 yang dapat dijabarkan sebagai berikut:

$$X=\begin{bmatrix} x_1^1 \\ x_2^1 \\ x_3^1 \\ x_4^1 \end{bmatrix}, C_1=\begin{bmatrix} x_1^2 \\ x_2^2 \\ x_3^2 \\ x_4^2 \end{bmatrix}, C_2=\begin{bmatrix} x_1^3 \\ x_2^3 \\ x_3^3 \\ x_4^3 \end{bmatrix},$$

sehingga matriks $A=\begin{bmatrix} x_1^1 & x_1^2 & x_1^3 \\ x_2^1 & x_2^2 & x_2^3 \\ x_3^1 & x_3^2 & x_3^3 \\ x_4^1 & x_4^2 & x_4^3 \end{bmatrix}$

Dimana vektor baris data hasil prediksi adalah X, dan vektor berkorelasi yang terpilih adalah C1 dan C2.

3) Tahap aproksimasi hasil prediksi: Tahap ini merupakan tahap utama dimana matriks A dari tahapan selanjutnya akan di proses menggunakan metode Centroid Decomposition. Proses aproksimasi akan berhenti jika iterasi telah mencapai jumlah iterasi tertentu atau telah mencapai ambang nilai yang telah ditentukan (threshold). Metode prediksi lain yang digunakan pada penelitian ini dan sebagai pembanding dari metode Centroid Decomposition adalah metode ekstrapolasi linear. Pada bidang matematika, proses ekstrapolasi sendiri merupakan sebuah proses menduga diluar rentang observasi aktual menggunakan nilai-nilai observasi yang ada, metode ini serupa dengan interpolasi yang dimana prosesnya adalah mencari nilai estimasi diantara nilai-nilai yang ada(Armstrong & Collopy 1993). Rumus dari ekstrapilasi linear dapat dilihat pada Pers 1.

$$y(x) = y_1 \frac{x-x_1}{x_2-x_1}(y_2 - y_1) \qquad (Pers.1)$$

C. *Evaluasi hasil*

Hasil-hasil yang didapat dari proses prediksi menggunakan metode-metode yang telah dijelaskan sebelumnya, selanjutnya akan di evaluasi seberapa akurat prediksi yang dihasilkan menggunakan metrik evaluasi RMSE (Root Mean Square Error). Persamaan RMSE dapat dilihat pada Pers. 2.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{mo\,del,i})^2}{n}} \qquad \text{(Pers.2)}$$

dimana,
n = jumlah data
Xobs = nilai data terobservasi/aktual
Xmodel = nilai data hasil dari model perhitungan

RMSE telah menjadi alat ukur dari hasil prediksi atau estimasi dan banyak digunakan oleh penelitian-penelitian sebelumnya(Lobato et al. 2015).

III. Hasil dan Pembahasan

Penelitian ini menggunakan sebuah metode yang memiliki pendekatan karakteristik yaitu Centroid Decomposition, sehingga hasil yang didapatkan dari metode tersebut akan menyerupai atau mendekati bentuk data secara umum.

Dataset yang digunakan dalam proses pengujian metode ini pada dasarnya telah melalui tahap rekayasa data untuk dapat menguji sejauh mana hasil yang didapat dari metode-metode yang digunakan terhadap nilai aktualnya.

Dari hasil evaluasi menggunakan RMSE hasil-hasil dari kedua metode terhadap nilai aktualnya dapat dilihat pada Tabel III.

Tabel IIIII
Perbandingan RMSE

Titik yang di Prediksi	**RMSE Ekstrapolasi Linear**	**RMSE *Centroid Decomposition***
2 Titik	1.339154	2.671527
5 Titik	5.738612	1.305149
10 Titik	14.17985	2.63317
15 Titik	2.257626	1.941969

Hasil pengujian menggunakan RMSE pada Tabel II menunjukan bahwa metode Ekstrapolasi Linear tidak memiliki konsistensi hasil saat jumlah titik prediksi semakin banyak. Pada 2 titik prediksi, RMSE dari metode Ekstrapolasi Linear lebih baik dibandingkan dengan metode Centroid

Decomposition. Akan tetapi, untuk 5 titik hingga 15 titik prediksi, metode Centroid Decomposition terbukti lebih baik daripada metode Ekstrapolasi Linear.

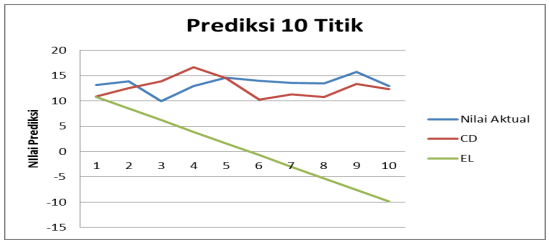
Hasil prediksi yang dihasilkan oleh metode Centroid Decomposition(CD) dan ekstrapolasi linear (EL) dapat dilihat pada Tabel IV.

TABEL IVV
HASIL PREDIKSI

Titik yang diprediksi	Nilai Aktual	CD	EL
2 titik	15.7	14.0881	13.4
	13	16.417	13.3
5 titik	14	14.96589	16.3
	13.6	13.13938	18
	13.5	12.53612	19.7
	15.7	14.03881	21.4
	13	14.9192	23.1
10 titik	13.2	10.85027	10.8
	13.9	12.56889	8.5
	10	13.90309	6.2
	12.9	16.69535	3.9
	14.6	14.546	1.6
	14	10.29142	-0.7
	13.6	11.30361	-3
	13.5	10.75764	-5.3
	15.7	13.37869	-7.6
	13	12.318	-9.9
15 titik	13.9	12.89563	13.8
	17.2	14.88816	14
	14.7	14.75333	14.2
	15.4	12.2401	14.4
	13.1	13.31669	14.6
	13.2	12.52286	14.8
	13.9	11.64984	15
	10	13.95175	15.2
	12.9	14.49452	15.4
	14.6	16.4476	15.6
	14	14.90435	15.8
	13.6	13.81345	16
	13.5	12.5777	16.2
	15.7	14.01087	16.4
	13	15.91874	16.6

Dari Tabel IV dapat dilihat bahwa nilai-nilai yang dihasilkan oleh ekstrapolasi linear pada awalnya terlihat baik-baik saja, namun ketika titik yang harus di prediksi semakin banyak, nilai yang dihasilkan akan menjadi sangat tidak konsisten. Sedangkan pada nilai prediksi yang dihasilkan oleh metode Centroid Decomposition nilainya cenderung lebih stabil walaupun tidak sepenuhnya mendekati nilai aktualnya.

Jika dilihat secara visual, maka gambaran hasil prediksinya dapat dilihat seperti pada Gambar 4.



Gbr. 4 Visualisasi Hasil Prediksi 10 Titik data

Dari Gambar 4 diatas dapat dilihat bahwa metode Ekstrapolasi Linear memiliki hasil yang bergantung pada trend umum dari data tersebut lalu dirunut secara linear sepanjang jumlah titik yang di prediksi.

Secara umum dilihat dari skenario pengujian yang telah dilakukan pada penelitian ini, metode Centroid Decomposition tetap menghasilkan neilai prediksi dengan stabil walaupun titik prediksinya bertambah. Namun hal tersebut tidak terjadi untuk metode Ekstrapolasi Linear, dimana seiring bertambahnya titik prediksi, hasilnya pun semakin memburuk. Maka dapat disimpulkan bahwa metode Ekstrapolasi Linear tidak cocok untuk melakukan prediksi jangka menengah hingga panjang, namun dapat dilakukan menggunakan metode Centroid Decomposition.

IV. SIMPULAN

Proses prediksi sebuah data dapat menentukan proses pengambilan keputusan terkait hal-hal yang dimasa datang, panjang pendeknya masa/interval waktu yang diprediksi juga dapat mempengaruhi seberapa jauh pengetahuan kita terhadap suatu hal. Metode-metode prediksi memiliki kekurangan dan kelebihannya masing-masing. Kegunaan dari masing-masing metode memiliki kecocokannya masing-masing pada kondisi-kondisi tertentu tetapi tidak dapat dikatakan baik untuk keseluruhan kondisi yang mungkin terjadi.

Penelitian-penelitian terkait prediksi cuaca telah banyak dilakukan oleh peneliti-peneliti terdahulu, dan hasil dari penelitian-penelitian tersebut tentunya menunjukan hasil yang lebih baik, meskipun tidak menggunakan jenis dataset yang serupa, sehingga tidak dapat dikatakan baik secara umum.

Dari penelitian ini, dapat diambils ebuah pengentahuan baru bahwa metode atau pendekatan yang dikembangkan untuk menanganai permasalahan Missing Values dan memiliki konsep dasar untuk melakukan Interpolasi data dapat digunakan untuk proses prediksi data yang memiliki konsep dasar Ekstrapolasi.

Penelitian ini menunjukan bahwa terdapat banyak sekali alternatif metode-metode yang dapat digunakan untuk melakukan proses prediksi dengan baik, meskipun metode tersebut secara pengembangan bukan dibuat untuk menangani masalah prediksi. Namun dikarenakan konsep Interpolasi dan konsep Ekstrapolasi tidak jauh berbeda, maka hasil dari metodenya pun terlihat baik.

Pada penelitian-penelitian kedepan, diharapkan dapat dilakukan pengujian-pengujian yang lebih komprehensif terkait metode Centroid Decomposition untuk menangani masalah prediksi data cuaca dengan melakukan pengujian komparasi dengan metode-metode prediksi yang lebih canggih.

DAFTAR PUSTAKA

[1] Armstrong, J.S. & Collopy, F., 1993. Causal Forces: Structuring Knowledge for Time-series Extrapolation. *JOURNAL OF FORECASTING*, 12, pp.103–115.

[2] Borzsonyi, E., 2013. Recovery of Missing Values based on Centroid Decomposition Eszter B o. University of Zurich.

[3] Han, J., Kamber, M. & Pei, J., 2011. *Data Mining: Concepts and Techniques* 3rd ed., San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[4] Lobato, F. et al., 2015. Multi-objective genetic algorithm for missing data imputation. *Pattern Recognition Letters*, 68, pp.126–131. Available at: http://dx.doi.org/10.1016/j.patrec.2015.08.023.

[5] Omary, A. et al., 2012. An interactive predictive system for weather forecasting. In *2012 International Conference on Computer, Information and Telecommunication Systems (CITS)*. pp. 1–4.

[6] Rahman, M.G. & Islam, M.Z., 2016. Missing Value Imputation Using a Fuzzy Clustering-based EM Approach. *Knowl. Inf. Syst.*, 46(2), pp.389–422. Available at: http://dx.doi.org/10.1007/s10115-015-0822-y.

[7] Shivaranjani, M.P., 2016. A Review of Weather Forecasting Using Data Mining Techniques. *International Journal Of Engineering And Computer Science*.