IMPLEMENTASI DAN ANALISIS METODE CLUSTERING-BASED MISSING VALUE IMPUTATION (CMI)

Imam Arbianto Wicaksono[1], Shaufiah[2], Imelda Ataina[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Missing value adalah kondisi pada data dimana data tersebut memiliki nilai atribut yang tidak lengkap atau hilang. Missing value pada data dapat ditangani dalam berbagai cara, salah satunya adalah imputasi missing value sehingga didapat nilai yang cocok untuk mengisi missing value. CMI (Cluster-based Missing Value Imputation) adalah salah satu metode missing value imputation pada data numerik yang berbasiskan klaster. CMI akan mengklasterkan data menggunkan K-means clustering kemudian akan mengisi missing value dengan bantuan fungsi kernel berdasarkan nilai-nilai yang ada pada klaster dimana missing value tersebut berada. Fungsi kernel pada CMI bertujuan untuk menangani non-parametric imputation dimana data tersebut dianggap tidak dapat diregresikan secara umum. CMI dapat menangani missing value dengan tingkat akurasi yang baik dilihat dari hasil RMSE pada pengujian langsung dan F-measure pada performansi klasifikasi pengujian tidak langsung yang didapatkan.

Kata Kunci : missing value, klaster, numerik, CMI, non-parametric

Abstract

Missing value is the condition on the data where the data has a value attribute that is incomplete or missing. Missing values in the data can be handled in different ways, one of which is the imputation of missing values so obtained are suitable for filling the missing value. CMI (Cluster-based Missing Value Imputation) is one of the missing value imputation methods of numerical data based on clusters. The data will be clustered by CMI use K-means clustering and then will fill in missing values with the help of kernel functions based on the values that exist in clusters in which the missing value is located. Kernel function in CMI aims to address non-parametric imputation where the data is not considered to be regressed in general. Tests performed were divided into two parts, namely the direct testing and indirect testing. Direct testing performed by calculating the RMSE of the estimates used to fill in missing values to actual values. Indirect testing is done by calculating F-measure of the classification results of test data on the actual class of data prior to testing. CMI can handle missing values with a relatively good level of accuracy seen from the results of RMSE and F-measure is obtained.

Keywords : missing value, cluster, numerical , CMI, non-parametric

1. Pendahuluan

1.1 Latar Belakang

Perkembangan teknologi informasi saat ini berkembang sangat pesat seiring kebutuhan manusia yang bertambah pula. Salah satu kebutuhan manusia yang perkembangan cukup pesat adalah kebutuhan akan informasi. Dikarenakan hal tersebut maka semakin hari semakin banyak data yang menumpuk. Untuk mengimbangi perkembangan kebutuhan manusia akan informasi, maka perkembangan teknologi informasi khususnya di bidang manajemen data turut dikembangkan. Terkadang data yang terkumpul belum memberikan informasi secara jelas, maka diperlukan metode yang menangani pencarian informasi dalam data. Metode yang dapat digunakan untuk memebrikan solusi masalah tersebut adalah metode *data mining*.

Ping-Ning Tan dan Kumar Steinbach M. [12] menyatakan bahwa d*ata mining* adalah sekumpulan proses otomatis untuk menggali informasi yang berguna dari data yang besar, sehingga didapatkan informasi yang berguna.*Data mining* adalah bagian dari *Knowledge Discovery in Database* (KDD), yang keseluruhan prosesnya bertujuan merubah data mentah menjadi informasi yang bermanfaat[12]. Proses-proses dari KDD adalah dimulai dari *input data*, lalu *data preprocessing*, *data mining*, *postprocessing*, kemudian terakhir menghasilkan informasi. Dalam proses *preprocessing* bertujuan untuk merubah data mentah menjadi format data yang dibutuhkan untuk proses data mining[12]. Proses *preprocessing* memiliki beberapa cara yang dapat dilakukan, diantaranya adalah *data cleaning, data integration, data transformation,* dan *reduction*. *Data cleaning* adalah proses menghilangkan noise atau data yang tidak relevan untuk proses data mining[7]. *Data cleaning* sendiri memiliki beberapa cara yaitu pengisian *missing value*, identifikasi *outlier*, *smooting noisy data*, dan *correct inconsistent data.*

Missing Value adalah adalah data yang memiliki nilai atribut yang tidak lengkap atau hilang[12]. Dalam beberapa kasus, hal ini terjadi karena *human-error* atau kesengajaan subjek untuk tidak mengisi nilai tertentu, misal beberapa orang tidak ingin mengisi kolom umur atau berat badan[12]. Namun bagaimanapun juga, *missing value* harus diikutsertakan dalam proses analisis data. Beberapa cara menangani *missing value* dapat menggunakan beberapa pendekatan diantaranya adalah :(a) tidak menganggap objek yang mengandung *missing value;* (b) mengisi *missing value* secara manual; (c) ganti nilai pada *missing value* dengan nilai yang global atau rata-rata nilai dari objek; (d) isikan *missing value* dengan nilai yang memiliki kemungkinan terbaik[14]. Pendekatan (a) memiliki kelemahan kemungkinan kehilangan informasi lebih banyak, untuk pendekatan (b) memakan waktu dan biaya, sedangkan pendekatan (c) mengganti semua *missing value* dengan nilai yang sama sehingga dapat menyebabkan gangguan dalam distribusi data[14]. Pendekatan (d) adalah pendekatan yang memungkinkan meminimalis kekurangan dari pendekatan-pendekatan sebelumnya, namun tetap tidak menutup kemungkinan akan terjadi kesalahan dalam proses penggalian informasi nantinya. Penanganan *missing value* penting dilakukan untuk data yang membutuhkan hasil akurasi yang baik saat penggalian informasi. Hal ini disebabkan *missing value* dapat menjadikan hasil penggalian informasi pada data tersebut menjadi tidak valid. Contoh data yang membutuhkan tingkat akurasi yang tinggi adalah data kesehatan.

Teknik imputasi *missing value* dapat dibagi menjadi dua yaitu imputasi regresi berparameter (*parametric regression imputation*) dan imputasi non-parameter (*non-parametric regression imputation*)[14]. *Data parametric* adalah data yang dapat diregresikan secara linear atau fungsi umum lainya, sedangkan *data non-parametric* adalah data yang tidak dapat diregresikan secara umum [14]. CMI (Clustering-based Missing Value Imputation) adalah metode imputation regresi non-parameter (*non-parametric regression imputation*) berbasis clustering untuk menangani *missing value* yang memanfaatkan K-means dan fungsi kernel [14]. Pada CMI dilakukan beberapa proses dalam penanganan *missing value*, pertama lakukan strategi proses clustering yang dilakukan dengan metode clustering K-means. Kedua, dilakukan strategi imputasi fungsi Kernel atau disebut juga regresi non-parameter kernel imputasi. Fungsi Kernel ini juga merupakan metode yang efektif menangani *missing value* untuk komputasi yang efisien, kuat, dan stabil [14]. Berikutnya dengan algoritma CMI, *missing value* dapat diisikan berdasarkan cluster yang terbentuk. Singkatnya, setelah seluruh data (termasuk missing value) dibagi menurut *cluste*r yang sesuai, maka setiap *missing value* dari data diisikan dengan nilai yang mungkin berdasarkan cluster masing-masing data tersebut. Dengan menggunakan algoritma CMI akan dapat menangani *missing value* data yang bersifat non-linear atau yang tidak dapat difungsikan secara umum, sehingga bias didapatkan hasil akurasi yang lebih baik saat proses mining data.

1.2 Perumusan Masalah

Dari latar belakang permasalahan maka masalah dapat dirumuskan sebagai berikut :

1. Bagaimana akurasi CMI dalam penanganan *missing value* pada data numerik berdasarkan parameter RMSE?
2. Bagaimana pengaruh penanganan *missing value* menggunakan CMI terhadap proses data mining (klasifikasi) berdasar parameter *precision, recall,* dan *F-measure*?

Batasan masalah dari penelitian Tugas Akhir ini yaitu :

1. Jenis data adalah data numerik karena dibutuhkan perhitungan matematis dalam proses fungsi kernel.
2. *Clustering* dilakukan dengan metode *K-means clustering*.
3. Klasifikasi dilakukan dengan metode C-45 pada aplikasi Weka 3.6.6.
4. Untuk evaluasi dengan klasifikasi, data yang digunakan tidak ada *missing value* pada atribut kelas.
5. Dalam data set yang digunakan tidak ada *missing value* atau bernilai *null* karena *missing value* akan dibentuk secara acak.
6. Format file untuk pada proses penanganan *missing value* dengan CMI adalah .mat.
7. Format file untuk proses klasifikasi pada Weka adalah .arff.

1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah sebagai berikut :

Mengimplementasikan dan mengetahui akurasi algoritma CMI dalam penanganan *missing value.*

1.4 Metodologi Penyelesaian Masalah

Metodologi yang digunakan dalam Tugas Akhir ini sebagai berikut :

1. Studi Literatur

   Pada tahap ini digunakan untuk mencari referensi dan data terkait data mining, missing value dan algoritma CMI. Referensi dapat berupa buku, website, paper, dan jurnal terkait yang akan dijadikan dasar teori.

2. Analisa kebutuhan dan perancangan perangkat lunak

   Tahapan untuk menentukan kebutuhan sistem, seperti identifikasi input, identifikasi output, identifikasi spesifikasi software dan hardware yang dibutuhkan.

   Dalam CMI, data input sebelumnya sudah melalui proses clustering menggunkan K-means, yang hasil dari clustering tersebut akan digunakan untuk lanjut ke proses berikutnya yaitu fungsi kernel. Setelah proses tersebut, maka dapat dilakukan missing value imputation. Setelah missing value ditangani, maka data dapat dilakukan proses mining sehingga diketahui akurasinya.

3. Implementasi perangkat Lunak

   Tahapan ini melakukan coding untuk pembangunan perangkat lunak penanganan *missing value* sekaligus analisis CMI algoritma yang diterapkan pada perangkat lunak.

4. Pengujian dan analisa perangkat lunak

   Tahap ini melakukan pengujian perangkat lunak dan analisis hasil pengujian. Pada tahap ini juga dilakukan evaluasi untuk mendapatkan hasil yang terbaik.

5. Pengambilan kesimpulan dan penyusunan laporan

   Tahap ini adalah tahap penarikan kesimpulan dari semua yang telah dilakukan dan penyelesaian penyusunan laporan. Penyusunan laporan dilakukan bertahap sejak awal pengerjaan Tugas Akhir.

5. Kesimpulan dan Saran

5.1 Kesimpulan

1. Jumlah klaster dan besar MVR berpengaruh pada hasil akurasi metode *Clustering-based Missing Value Imputation* (CMI) dalam penanganan *missing value*.
2. Besar variansi nilai atribut data dan nilai kelas berpengaruh pada hasil akurasi metode *Clustering-based Missing Value Imputation* (CMI) dalam penanganan *missing value.*
3. Metode *Clustering-based Missing Value Imputation* (CMI) dapat menangani *missing value* dengan performansi cukup baik.
4. Metode *Clustering-based Missing Value Imputation* (CMI) lebih cocok digunakan untuk data numerik yang memiliki variansi nilai atribut dan nilai kelas yang kecil.

5.2 Saran

1. Untuk metode imputasi *missing value* yang berbasis *clustering* dapat dicoba menggunakan metode *clustering* yang lain sehingga dapat membanding metode *clustering* mana yang lebih baik.
2. Dapat dicoba mengembangkan metode CMI untuk data selain data numerik.

Daftar Pustaka

[1] Agusta, Yudi. 2011 . "K-Means" . wordpress.com . http://yudiagusta.wordpress.com/k-means [Diakses tanggal 25 Maret 2011]

[2] Anbasari, M.S., Mehata, K.M. 2010 . "Enhanced K-Means Clustering for Patient Reported Outcome". Department of Computer Science and Engineering, Anna University , India

[3] Cindy C.S., Dian . 2009 . "Analisis Metode K-Means Imputation untuk Penanganan Missing Value" . Bandung : Fakultas Informatika, Institut Teknologi Telkom

[4] Denton, Anne and Perrizo, William. 2008 . "A Kernel-Based Semi-Naive Bayesian Classifier Using P-Tress". Departement of Computer Science, North Dakota State University, USA

[5] Do, Thanh-Nghi and Poulet, Francois. 2005 . "Kernel Method and Visualization for Interval Data Mining". College of Information Technology, Can Tho University, Vietnam and ESIFA Pole ECD, Laval- France

[6] Han, Jiawei , Kamber, Micheline . 2006. "Data Mining : Concepts and Techniques". 2nd ed. San Francisco : Morgan Kaufmann

[7] Huda, Nuqson Masykur . 2010 . "Aplikasi Data Mining Untuk Menampilkan Tingkat Kelulusan Mahasiswa". Semarang : Program Studi Teknik Informatika, Jurusan Matematika, Fakultas Matematika dan IPA, Universitas Diponegoro

[8] Moertini, Veronika S. 2002. " Data Mining Sebagai Solusi Bisnis". Integral , vol. 7 no.1

[9] Pressman, Roger S . 1997 . "Software Engineering : A Practitioner's Approach " . Diterjemahkan dalam Bahasa Indonesia oleh L.N. Harnaningrum . 2002 . Yogyakarta : Penerbit Andi

[10] Rey-del-Castilo, Pilar and Cardenosa, Jesus . 2009 . "Catagorical Missing Data Imputation Using Fuzzy Neural Networks with Numerical and Catagorical Inputs". World Academy of Science, Engineering and Technology, 55, halaman 628-635

[11] Sitohang, Benhard , Saptawati , G.A. Putri .2006. "TPDA ^{2}Algorithm for Learning BN Structure From Missing Value And Outlier in Data Mining". Jurnal Informatika , Vol. 7 no. 2 , halaman 108-113

[12] Tan, Ping-Ning , Steinbach M. , Kumar , V. 2006. "Intoduction To Data Mining". Boston : Pearson Education, Inc.

[13] Zhang, Chengqi , Qin, Yongsong , Zhu , Xiaofeng , Zhang, Jilian , Zhang, Shica .2006. "Clustering-based Missing Value Imputation for Data Preprocessing". University of Technology Sydney. Australia

[14] Zhang, Shicao , Zhang, Jilian , Qin, Yongsong , Zhang, Chengqi . 2007 . "Missing Value Imputation Based On Data Clustering" . Department Of Computer Science, Guangxi Normal University, Guilin, China