ANALISIS METODE K-MEANS IMPUTATION UNTUK PENANGANAN MISSING VALUE

Dian Cindy Sirait[1], Arie Ardiyanti Suryani[2], Angelina Prima Kurniati[3]

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

Abstrak

Missing value merupakan salah satu penyebab suatu data dikatakan tidak valid atau masih merupakan data 'kotor'. Missing value dengan jumlah banyak akan menghilangkan sejumlah informasi yang seharusnya dibutuhkan untuk diolah. Salah satu metode penanganan yang dapat dilakukan pada missing value adalah dengan mengisikan suatu nilai kepada data-data yang missing. Nilai untuk mengisi missing value diperoleh dengan prediksi dari informasi yang masih tersedia pada data. Metode ini disebut dengan metode imputasi. Salah satu metode imputasi yaitu K-Means Imputation (KMI). Metode ini adalah metode yang didasarkan pada algoritma K-Means yang biasa dipakai dalam proses clustering. KMI akan melakukan pengisian missing value dengan menghitung rata-rata nilai atribut dari data-data yang berada pada suatu cluster dimana data missing juga berada dan data-data tersebut haruslah memiliki kelas yang sama dengan yang dimiliki data missing. Pada tugas akhir ini, akan dicoba untuk melakukan imputasi missing value dengan menggunakan K-Means Imputation (KMI). Parameter yang akan diuji adalah root mean square error, precision, recall dan f-measure. Berdasarkan hasil pengujian, KMI terbukti dapat memprediksi nilai missing value mendekati nilai data yang sebenarnya pada data yang memiliki tingkat keragaman variasi nilai atribut kecil dan dapat meningkatkan hasil klasifikasi.

Kata Kunci : metode imputasi, K-Means Imputation (KMI), missing value.

Abstract

Missing value is one of causes that make the data is called invalid data. Data with many missing value, will take the necessary information is missing. One of the way to handle missing value is fill the missing data with the predict value from available information in data. This method is called imputation method. One of the imputation methods is K-Means Imputation (KMI). This algorithm based on K-Means algorithm that is usually used in clustering. KMI will fill missing value with calculate mean atribute value from datas in a cluster where data missing is belong and those datas have to have class atribute as same as with class missing value data. In this final project, it will try to imputate missing value with K-Means Imputation (KMI). Some parameters that will be tested are root mean square error, precision, recall and f-measure. Refers to the result of the experiment, KMI is proved predicting value approach to the real value for data which consist of few variation atributte value and it can make the classification result be better.

Keywords : imputation method, K-Means Imputation (KMI), missing value.

1. Pendahuluan

1. 1 Latar belakang masalah

Data yang ada di dunia nyata biasanya masih berupa ‘data kotor’ atau memiliki beberapa masalah seperti adanya missing value, data duplikat, dan *inconsistent data*. Salah satu masalah yang sering dihadapi adalah adanya *missing value* pada data yang akan analisis. *Missing value* merupakan data atau informasi yang hilang atau tidak tersedia pada variabel tertentu akibat beberapa faktor seperti kesalahan prosedur saat *entry* data, penyimpanan data yang kurang baik, dan kemungkinan atribut tersebut tidak relevan pada semua kasus misalnya atribut pendapatan tahunan tidak dipakai untuk anak-anak dan berbagai sebab lain yang memungkinkan adanya ketidaktersediaan informasi secara lengkap. *Missing value* merupakan masalah yang tampaknya sepele dan sering dijumpai pada dataset di dunia nyata. Tidak heran beberapa kalangan membiarkan hal itu begitu saja. Padahal dengan adanya missing value pada data yang akan diolah akan mengurangi tingkat keakuratan data tersebut saat akan diolah lebih jauh, apakah akan diklasifikasi, diklasterisasi, atau diasosiasi. Untuk itulah, diperlukan suatu penanganan khusus terhadap masalah kemunculan missing value.

Ada beberapa cara yang dapat dilakukan untuk penanganan missing value diantaranya menghapus baris data yang memiliki missing value, dan mengisikan missing value dengan nilai estimasi dari informasi data yang tersedia. Pada Tugas Akhir ini diperkenalkan sebuah metode pengisian missing value yaitu metode imputasi dimana prediksi nilai missing value dilakukan dengan menggunakan algoritma tertentu yaitu salah satunya adalah algoritma *K-Means Imputation*(KMI).

KMI akan melakukan *clustering* terhadap objek-objek data termasuk missing value ke dalam *cluster-cluster* dengan melakukan perhitungan jarak tiap objek data terhadap centroid yang ada. Tiap cluster terdiri dari objek data yang memiliki karakteristik yang sama yaitu memiliki jarak minimum terhadap centroid di cluster tersebut. *Missing value* dapat diisikan setelah proses iterasi pembentukan k-cluster berhenti dimana didapatkan keanggotaan di suatu cluster yang stabil atau tidak ada perubahan anggota objek data di cluster tersebut. Nilai yang akan diisikan pada missing value adalah nilai rata-rata atribut objek data disuatu cluster dari atribut kelas dimana data missing value berada.

1.2 Perumusan masalah

Permasalahan yang dijadikan objek penelitian dan pengembangan tugas akhir ini adalah sebagai berikut :

- Bagaimana melakukan penanganan data yang memiliki missing value dengan menggunakan metode K-Means Imputation.
- Bagaimana menganalisis performansi dari metode K-Means Imputation berdasarkan parameter-parameter tertentu seperti *Root Mean Square Error* (RMSE), *Precision*, *Recall* dan *F-Measure*.

1.3 Tujuan

Dalam Tugas Akhir ini , tujuan yang ingin dicapai antara lain :

- Menganalisis metode K-Means Imputation untuk penanganan missing value dengan missing value rate yang berbeda.
- Menganalisis sejauh mana pengaruh performansi terhadap jumlah cluster yang berbeda berdasarkan *Root Mean Square Error* (RMSE).
- Menganalisis performansi penanganan missing value dengan metode K–Means Imputation menggunakan metode classifier yang ada pada Weka dalam proses klasifikasi berdasarkan *Precision*, *Recall* dan *F-Measure*.

1.4 Batasan Masalah

Masalah yang akan dibahas memiliki batasan-batasan sebagai berikut:

- Data yang digunakan adalah data siap pakai dalam bentuk tabel yang berisi distribusi atribut dan value atribut.
- Tipe data yang digunakan adalah data bertipe numerik.
- Data yang digunakan tidak memiliki data error dalam distribusi instancenya.

1.5 Metodologi penyelesaian masalah

Metode yang akan digunakan untuk menyelesaikan tugas akhir ini adalah :

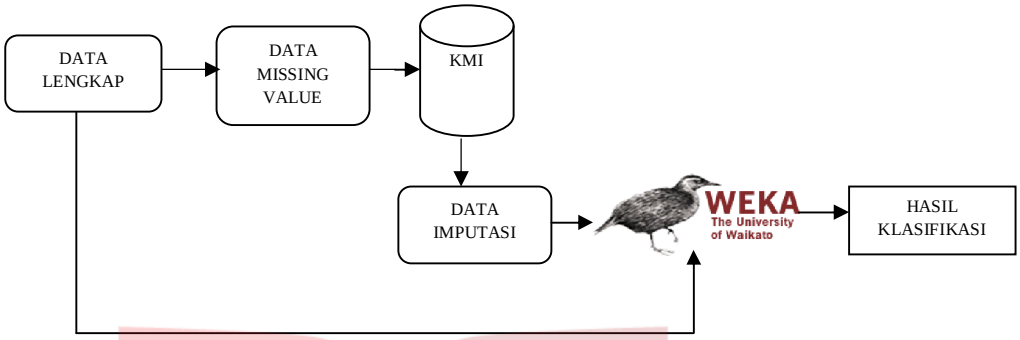
- Studi literatur dan pengumpulan data

  Mempelajari landasan teori dari literatur-literatur (jurnal, paper, artikel dan buku) yang berkaitan dengan topik Tugas Akhir dan mengumpulkan dataset yang nantinya dipakai saat pengujian yang diperoleh dari internet.
- Analisis dan Perancangan

  Menganalisis dan merancang penerapan metode imputasi pada data yang memiliki missing value dengan menerapkan algoritma *K-Means Imputation* (KMI).
- Implementasi perancangan perangkat lunak

  Implementasi perancangan dengan membangun suatu perangkat lunak yang menerapkan KMI.
- Analisis dan evaluasi

  Memproses data melalui perangkat lunak yang dibuat kemudian menganalisis performansi hasil yang diperoleh dengan melihat pengaruh penanganan missing value menggunakan KMI terhadap akurasi klasifikasi data. Parameter pengujian yang dipakai yaitu *Root Mean Square Error* (RMSE), *Precision*, *Recall* dan *F-Measure*.



Gambar 1.1 Tahapan Skenario Analisis dan Evaluasi

- Pembuatan Laporan

  Mendokumentasikan hasil pembuatan sistem dari mulai tahap studi literatur sampai tahap pengambilan kesimpulan ke dalam suatu bentuk laporan.

5. Penutup

5.1 Kesimpulan

Berdasarkan analisis terhadap hasil pengujian perangkat lunak dalam Tugas Akhir ini dapat dihasilkan kesimpulan sebagai berikut:

1. Hasil *Root Mean Square Error* (RMSE) dipengaruhi banyaknya cluster yang terbentuk. Jumlah cluster yang bisa terbentuk dipengaruhi jumlah ketersediaan *record complete*. Sedangkan jumlah *record complete* ditentukan dari jumlah *instance* data, jumlah atribut, *Missing Value Rate*(MVR), dan *missing random*.
2. Variasi data pada atribut-atribut dalam dataset juga menentukan hasil RMSE. Semakin banyak variasi nilai yang dimiliki atribut, maka semakin besar RMSE yang diperoleh. Dengan kata lain keakuratan nilai prediksi untuk *missing value* semakin berkurang.
3. Semakin mirip nilai prediksi pada pengisian missing value dengan data lengkap, maka nilai *precision*, *recall* dan *f-measure* yang diperoleh semakin mendekati data yang lengkap atau data sebenarnya.
4. Dari pengujian yang telah dilakukan, dapat disimpulkan bahwa KMI memiliki nilai RMSE yang baik untuk data dengan variasi nilai atribut kecil.
5. Data imputasi dengan jumlah cluster yang berbeda tidak memiliki perubahan pengaruh yang besar pada saat dilakukan klasifikasi. Sedangkan dari pengukuran RMSE menunjukkan jumlah cluster yang berbeda memberikan perbedaan hasil RMSE yang tidak signifikan.

5.2 Saran

Beberapa saran yang dapat diperhatikan untuk mengembangkan Tugas Akhir ini:

1. Data yang digunakan tidak hanya file berformat *.xls.
2. Mencoba alternatif metode imputasi lain yang lebih baik dari segi keakuratan melakukan imputasi missing value.

Daftar Pustaka

[1]. Acuna, E., Caroline Rodriguez. *The Treatment of Missing Values and its Effect in the Classifier Accuracy.* Puerto Rico : Department of Mathematics, University of Puerto Rico. Diunduh pada tanggal 21 Juni 2009.

[2]. Agusta, Y., *K-Means* – Penerapan, Permasalahan dan Metode Terkait. *Jurnal Sistem dan Informatika Vol. 3 (Pebruari 2007), 47-60*. Diunduh pada tanggal 21 Juni 2009.

[3]. Busse,Jerzy W. Grzymala, dan Ming Hu, 2001. *A Comparison of Several Approaches to Missing Attribute Values in Data Mining*. W. Ziarko and Y. Yao (Eds.): RSCTC 2000, LNAI 2005, pp. 378−385. Diunduh pada tanggal 8 Juni 2009.

[4]. Farhangfara, A., Lukasz Kurgan, Jennifer Dy. 2008. *Impact of imputation of missing values on classification error for discrete data.* Pattern Recognition 41 (2008) 3692 – 3705. Diunduh pada tanggal 21 Juni 2009

[5]. Fayyad U., Gregory Shapiro, dan Padhraic Smyth. *From Data Mining to Knowledge Discovery in Database*. http://www.daedalus.es/fileadmin/daedalus/doc/MineriaDeDatos/fayyad96.pdf. Diunduh tanggal 16 Juli 2009.

[6]. *Imputation*. *http://www.statcan.gc.ca/edu/power-pouvoir/ch3/5214783-eng.htm*. diunduh pada tanggal 21 Juni 2009.

[7]. Li, D., Jitender Deogun, William Spaulding dan Bill Shuart, 2004. *Towards Missing Data Imputation: A Study of Fuzzy K-Means Clustering Method*. S.Tsumoto ct al. (Eds): RSCTC 2004, LNAI 3066,pp. 573-579. Diunduh pada tanggal 8 Juni 2009.

[8]. Mehala, B., K. Vivekanandan dan P.Ranjit Jeba Thangaiah, 2008. *An analysis on K-Means Algorithm as an Imputation Method to Deal with Missing Values*. Asian Journal of Information Technology 7(9): 434-441. Diunduh pada tanggal 8 Juni 2009.

[9]. *Research Digest* : Missing diisi tak berarti manipulasi. http://researchexpert.wordpress.com. Diunduh pada tanggal 21 Juni 2009

[10]. Wagstaff, K. *Clustering with Missing Value*s: *No Imputation Required*. California: California Institute of Technology. Diunduh pada tanggal 8 Juni 2009.

[11]. Wagstaff, K., dan Victoria G. Laidler,2005. *Making the Most of Missing Value Object Clustering with Partial Data in Astronomy*. Astronomical Data Analysis Software and Systems XIV ASP Conference Series Vol. XXX. Diunduh pada tanggal 8 Juni 2009.

[12]. Zhang, Shichao, Jilian Zhang, Xiaofeng Zhu, Yongsong Qin dan Chengqi Zhang. *Missing Value Imputation Based on Data Clustering*. Diunduh pada tanggal 8 Juni 2009.