IMPUTASI *MISSING ATTRIBUTE VALUES DATASET* HEPATITIS BERDASARKAN ALGORITME *RIPPER*

Oleh :
Tri Astuti[1], Yuliyanti[2]

Teknik Informatika[1], Sistem Informasi[2]
STIMIK AMIKOM Purwokerto
tri_astuti@amikompurwokerto.ac.id[1], yooli_juli@yahoo.com[2]

Abstract

Hepatitis is a liver disease which caused by a hepatitis virus. Nowdays hepatitis is a global health problems, including in Indonesia. Chronic hepatitis can lead to cirrhosis and liver cancer, therefore early diagnosis is needed. The diagnosis process of hepatitis disease are done through computer aided method using hepatitis dataset nowdays. University California Irvine (UCI) machine learning repository has been providing hepatitis disease dataset which can be accessed to public but the dataset contains many missing values. The existing of missing values in the dataset may affect the quality of the analysis results, therefore it needs to be conducted for handling the missing values. Imputation method based on machine learning is one of the methods to handle the missing value. The aims of this research is to develop the imputation methods of missing value using machine learning algorithm based on RIPPER on hepatitis dataset. Result shows that the imputation method based on RIPPER achives 87,50% accuracy for hepatitis dataset. It is expected that the developed method can contribute for helping the clinicans and practicians by providing imputed hepatitis dataset in diagnosing the hepatitis disease.

Keywords : Hepatitis, missing values, imputation

A. PENDAHULUAN

Hati merupakan kelenjar terbesar dan organ yang penting bagi tubuh. Hati berfungsi sebagai alat ekskresi dan proses detoksifikasi. Hati memproduksi protein dan gula untuk dimanfaatkan tubuh. Vitamin, gula, lemak dan nutrisi yang tersimpan dalam hati akan dilepaskan pada saat tubuh membutuhkannya (Grabowski, 2011). Hati bekerjasama dengan ginjal memecah beberapa senyawa yang bersifat racun dan menghasilkan amonia, urea, dan asam urat dengan memanfaatkan nitrogen dari asam amino (Anonymous,2014). Penyakit hati *(liver)*

dapat disebabkan oleh beberapa hal antara lain alkohol, obat-obatan, penyakit genetik berupa kelainan membran sel darah merah *(hereditary)* atau virus.

Penyakit hepatitis merupakan jenis penyakit hati yang disebabkan oleh virus. Virus tersebut meliputi virus hepatitis A, hepatitis B, hepatitis C, hepatitis D dan hepatitis E. Penyakit hepatitis merupakan permasalah kesehatan global, termasuk di Indonesia. Prevalensi infeksi virus hepatitis B (VHB) berbeda-beda di seluruh dunia. Angka prevalensi penderita hepatitis di tingkat Asia Pasifik cukup tinggi mencapai 8% dengan penularan secara vertikal dan horizontal (Kemenkes, 2013).

Diantara negara-negara anggota World Health Organisation South East Asian Region (WHO SEAR), Indonesia merupakan negara dengan pengidap Hepatitis B nomor 2 terbesar sesudah Myanmar. Jenis hepatitis yang banyak menginfeksi penduduk Indonesia adalah hepatitis B (21,8%) dan hepatitis A (19,3%). Sekitar 23 juta penduduk Indonesia telah terinfeksi hepatitis B dan 2 juta orang terinfeksi hepatitis C [3]. Penyakit hepatitis A sering muncul dalam bentuk kejadian luar biasa (KLB) seperti yang terjadi di beberapa tempat di Indonesia. World Health Assembly (WHA) mengadopsi resolusi tentang virus hepatitis, menyerukan semua negara anggota WHO untuk melaksanakan pencegahan dan penanggulangan virus hepatitis secara komprehensif.

Seiring dengan perkembangan teknologi informasi diagnosis penyakit hepatitis banyak dikembangkan menggunakan metode computer aided. Metode diagnosis *computer aided* ini banyak memanfaatkan algoritme *machine learning* menggunakan *dataset* pasien pengidap penyakit hepatitis. Pendeteksian dini penderita penyakit hepatitis adalah kunci mencegah penyakit hepatitis kronis dan kematian yang diakibatkan oleh hepatitis.

Diagnosis berbasis komputer terkendala dengan ketersediaan dataset penyakit hepatitis yang lengkap. Repositori UCI telah menyediakan dataset penyakit hepatitis yang dapat diakses secara umum akan tetapi dataset tersebut banyak mengandung *missing values* (UCI Repository, 2014). *Missing values* adalah informasi yang tidak tersedia untuk sebuah objek atau nilai yang hilang dalam suatu dataset. Missing values pada dasarnya tidak bermasalah bagi keseluruhan data apabila jumlahnya hanya sedikit, misal hanya 1% dari seluruh

data. Namun jika persentase data yang hilang tersebut cukup besar, maka perlu dipertimbangkan suatu teknik untuk mengestimasi nilai pengganti missing values tersebut karena dapat mempengaruhi hasil uji suatu data. Pengolahan data akan menemui kesulitan dan validitas hasil analisis dipertanyakan jika pada dataset terdapat *missing values* (Bergmeir dan Benitez, 2012)(Aydilek dan Arslan, 2010)(Genc dkk., 2010)(Hulse dan Khoshgoftaar, 2007)(Nelwamondo dkk, 2013)(Acuna dan Rodriguez, 2004). Hal tersebut dikarenakan prasangka dan penurunan kualitas data dan algoritme (Yongsong dkk, 2007).

Fokus makalah ini adalah pada penanganan *missing values* pada *dataset* hepatitis dari repositori UCI *Machine Learning* berdasarkan algoritme RIPPER. Tulisan ini selanjutnya menjelaskan tentang *missing values* dan RIPPER pada bagian kedua. Bagian ketiga tentang metode imputasi yang diterapkan dan dilanjutkan dengan bab empat menjelaskan tentang percobaan yang dilakukan dan hasil. Bagian terakhir menjelasan kesimpulan dan *future study.*

B. TINJAUAN PUSTAKA

1. *Missing Values*

Missing values merupakan kondisi yang tidan diinginkan dalam *data mining*, *machine learning* dan sistem informasi (Aydilek dan Arslan. 2013). Hal ini juga merupakan permasalahan yang umum dalam analisis di bidang medis. Sejumlah metode imputasi telah diusulkan oleh peneliti terdahulu antara lain Izzah dkk. (2013) menggunakan algoritma pengelompokan data k-*harmonic means.*

Missing data dikategorikan menjadi tiga kategori (Kemenkes, 2013)(Thangavel dan Pethalakshmi, 2009). Kategori tersebut meliputi *Missing Completely at Random* (MCAR) jika data *missing* tidak ada relasi dengan variabel yang lain. *Missing at Random* (MAR) apabila data yang hilang memiliki hubungan dengan fitur yang lain. *Missing Not at Random* (MNAT) in adalah *missing values* yang memiliki relasi dengan data missing yang lain dan tidak dapat dilakukan estimasi dengan menggunakan variabel yang ada.

Metode imputasi dikelompokkan menjadi manual imputation, global constanta imputation, conventional imputation dan model imputasi prediksi (Chu

dkk., 2010). Sekarang ini algoritme *machine learning* telah dikenalkan dalam pengembangan metode imputasi (seftyawan dkk., 2013)(Pous dkk., 2008).

2. *RIPPER*

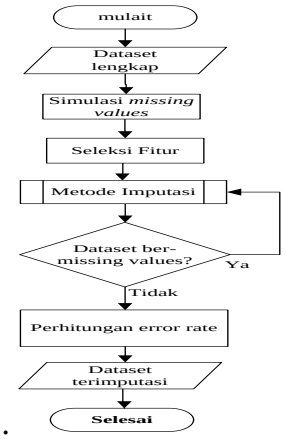
Algoritme RIPPER, *Repeated Incremental Prunning to Error Reduction* yaitu algoritme *machine learning* berbasis *rule* yang merupakan penyempurnaan dari *Incremental Reduce Error Prunning* (IREP) (Cohen, 1995).

Ripper mempertimbangkan dua alternatif pada *rule*, pertama disebut penggantian *rule,* mulai dari *rule* yang kosong, ditumbuhkan dan di-*prune.* Kedua disebut revisi *rule*, mulai dengan *rule* itu sendiri, ditumbuhkan dan kemudian di-*prune.* Dua metode ini dibandingkan dengan rule original, dan tiga terpendek ditambahkan pada *rule base*. Optimisasi *rule base* dapat dilakukan sebanyak kali, biasanya dua kali (Alpaydin, 2010).

C. METODE PENELITIAN

Pada tulisan ini, imputasi nilai-nilai yang hilang dilakukan sebagai berikut. Pertama, dataset lengkap ditetapkan sebagai dataset masukan dalam proses imputasi. Nilai-nilai yang hilang diciptakan secara buatan. Setelah nilai-nilai yang hilang dibuat, proses imputasi dilakukan melalui metode yang diusulkan. Kinerja imputasi missing values diukur dengan mencocokkan nilai-nilai diperhitungkan dengan nilai riil dalam atribut yang hilang. Metode evaluasi yang digunakan untuk mengevaluasi kinerja dari metode imputasi dengan menghitung tingkat kesalahan (Othman dkk., 2010)(Martono dkk., 2012).

Diagram alir penelitian yang mengambarkan metode yang diusulkan ditunjukkan pada Gambar 1. Proses imputasi dilakukan dengan *dataset* lengkap sebagai masukan awal, simulasi *missing values*, selanjutnya dilakukan proses imputasi menggunakan metode yang diusulkan kemudian menghitung tingkat kesalahan.



Gambar 1. Diagram alir metode yang diusulkan

Tingkat kesalahan proses imputasi dirumuskan pada persamaan sebagai berikut.

$$Errorrate = \frac{n\_number\_of\_incorrect\_prediction}{Total\_number\_of\_prediction} \quad (1)$$

D. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan *dataset* hepatitis UCI untuk proses imputasi *missing values* (Grawbowski, 2011). *Dataset* ini terdiri dari 155 kasus dengan 19 kondisi atribut dan satu atribut keputusan. Daftar atribut pada *dataset* hepatitis dapat dilihat pada Tabel I. *Dataset* ini memiliki dua kelas keputusan, *die* dan *live*. Penelitian ini dilakukan untuk memperkirakan atribut malaise yang telah disimulasikan ber-*missing values* secara artifisial. Nilai atribut ini dibagi menjadi kelas *yes* dan *no*.

Preprocessing dilakukan dengan membuat *dataset* lengkap merupakan proses pertama kali dilakukan dalam imputasi. *Dataset* lengkap ini diperoleh dengan menghapus contoh yang memiliki satu atau lebih nilai yang hilang dalam fitur *dataset*. Setelah menghapus semua nilai yang hilang kemudian mendapat 80 *instance* yang lengkap dan digunakan dalam proses imputasi hilang nilai. Kemudian dataset dibagi menjadi data pengujian dan pelatihan dengan membagi secara *stratified*. Evaluasi dilakukan dengan *k-fold cross validation*. Penelitian ini menggunakan *percentage split* 70:30 untuk membagi *training* dan *testing* data.

Tabel I. Daftar atribut hepatitis *dataset*

No	**Atribut**	**Nilai Atribut**
1	*Age*	10, 20, 30, 40, 50, 60, 70, 80
2	*Sex*	*male, female*
3	*Steroid*	*no, yes*
4	*Antivirals*	*no, yes*
5	*Fatigue*	*no, yes*
6	*Malaise*	*no, yes*
7	*Anorexia*	*no, yes*
8	*Liver big*	*no, yes*
9	*Liver firm*	*no, yes*
10	*Spleen palpable*	*no, yes*
11	*Spiders*	*no, yes*
12	*Ascites*	*no, yes*
13	*Varices*	*no, yes*
14	*Bilirubin*	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
15	*Alk phosphate*	33, 80, 120, 160, 200, 250
16	*Sgot*	13, 100, 200, 300, 400, 500
17	*Albumin*	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
18	*Protime*	10, 20, 30, 40, 50, 60, 70, 80, 90
19	*Histology*	*no, yes*
20	*Class*	*die, live*

Algoritme RIPPER digunakan sebagai metode imputasi untuk memperkirakan nilai yang hilang dalam atribut malaise. *Dataset* lengkap digunakan sebagai data pelatihan dan data pengujian memiliki nilai yang hilang dalam malaise atribut. Hasil dari metode imputasi berdasarkan algoritme RIPPER ditunjukkan pada Tabel II.

Tabel I. Performa Metode Imputasi

Method	**RIPPER**
Error rate	12.50 %

Berdasarkan Tabel II diketahui bahwa dengan memanfaatkan seluruh fitur pada *dataset* hepatitis untuk proses imputasi *missing values* attribute *malaise* diperoleh *error rate* yang diperoleh sebesar 12,50% atau akurasi sebesar 87,50%.

E. KESIMPULAN DAN SARAN

Tujuan utama dari penelitian ini adalah untuk mengetahui kinerja metode imputasi berdasarkan algoritme RIPPER. Hasil menunjukkan performa metode imputasi sebesar 87,50% pada pengujian menggunakan *percentage split test*. Perbandingan jumlah *training* data dan *testing* data sebesar 70:30. *Future studi* disarankan penggunaan dataset yang lebih besar untuk mendapatkan kinerja yang lebih representatif. Pengujian juga dapat diujicobkan menggunakan *k-fold cross validation* untuk mengurangi nilai bias dari akurasi yang diperoleh. Berbagai variasi nilai k pada *k-fold cross validation* juga disarankan untuk memperoleh pemahaman metode imputasi yang lebih baik dapat diperoleh.

DAFTAR PUSTAKA

Anonymous. 2014, "Penyakit Hepatitis A, Hepatitis B, Hepatitis C." [Online]. Available: http://penyakithepatitis.org/. [Accessed: 25-Jun-2014].

Acuna,E. and Rodriguez,C., 2004. "The Treatment of Missing Values and its Effect in the Classifier Accuracy," presented at the In Banks,D. et al. (eds) Classification, Clustering and Data Mining Applications, Springer-Verlag, Berlin, Heidelberg, pp. 639–648.

Aydilek ,I. B. and A. Arslan, 2013. "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," *Inf. Sci.*, vol. 233, pp. 25–35, Jun.

Bergmeir,C. and J. M. Benítez, 2012 "On the use of cross-validation for time series predictor evaluation," *Inf. Sci.*, vol. 191, pp. 192–213.

Chu, N., L. Ma, J. Li, P. Liu, and Y. Zhou, 2010 "Rough set based feature selection for improved differentiation of traditional Chinese medical data," in *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, vol. 6, pp. 2667–2672.

Grabowski,Chester. 2011 "Chronic Hepatitis B Virus Infection," presented at the Hepatitis B Health Conference, Amerika Serikat.

Genç, S. , F. E. Boran, D. Akay, and Z. Xu, 2010."Interval multiplicative transitivity for consistency, missing values and priority weights of interval fuzzy preference relations," *Inf. Sci.*, vol. 180, no. 24, pp. 4877–4891, Dec.

Hulse, J. Van and T. M. Khoshgoftaar, 2007. "Incomplete-Case Nearest Neighbor Imputation in Software Measurement Data," in *IEEE International Conference on Information Reuse and Integration, 2007. IRI 2007*, pp. 630–637.

Kemenkes RI. 2013.Badan Penelitian dan Pengembangan Kesehatan Republik Indonesia, "Riset Kesehatan Dasar 2013,".

Martono G. Hendro, Teguh Bharata Adji, and N. A. Setiawan, 2012 "Penggunaan Metodologi Analisa Komponen Utama (PCA) untuk Mereduksi Faktor-Faktor yang Mempengaruhi Penyakit Jantung Koroner," presented at the National Conference of Science, Engineering and Technology, 2012.

Nelwamondo, F. V. , D. Golding, and T. Marwala, 2013."A dynamic programming approach to missing data estimation using neural networks," *Inf. Sci.*, vol. 237, pp. 49–58, Jul.

Othman, N. B. O. M. S. Bin , F. Binti Jusoh, N. Binti Omar, and R. Binti Ibrahim, 2010 "Review of Feture Selection for solving Classification Problem," *J. Inf. Syst. Res. Innov.*,vol. 3, pp. 64–70.

Pous, C. , D. Caballero, and B. Lopez, 2008 "Diagnosing Patients Combining Principal Components Analysis and Case Based Reasoning," in *Eighth International Conference on Hybrid Intelligent Systems, 2008. HIS '08*, pp. 819–824.

Seftyawan,A. Itsnaini, D. Eka Ratnawati, and L. Muflikhah, 2013 "Penanganan Missing Value dengan Algoritma Weighted KNNI Pada Data Kategori," *DORO Repos. J. Mhs. PTIIK Univ. Brawijaya*, vol. 1 No.2.

Thangavel, K. and A. Pethalakshmi, 2009 "Dimensionality reduction based on rough set theory: A review," *Appl. Soft Comput.*, vol. 9, no. 1, pp. 1–12.

UCI Repository. 2014. "Dataset UCI Machine Learning," *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml/datasets.html. [Accessed: 08-Jun-2014].

Yongsong, Q. , Z. Shichao, Z. Xiaofeng, Z. Jilian, and Z. Chengqi, 2007."Semi-parametric optimization for missing data imputation," *Appl. Intell.*, vol. 27(1), pp. 79–88., Jan.