ABSTRAK

Dataset dimensi tinggi secara umum memiliki permasalahan yaitu adanya nilai yang hilang, yang disebabkan oleh kesalahan *human error*. Kesalahan tersebut terjadi karena pengamatan data tidak berjalan dengan baik, pengukuran tidak lengkap, dan permasalahan responden menolak untuk menjawab beberapa pertanyaan tertentu dalam survei. Nilai yang hilang dianggap bermasalah, karena nilai tersebut memiliki pengaruh besar pada saat pengambilan keputusan. Teknik imputasi digunakan untuk menghasilkan nilai pengganti yang hilang, sehingga diperoleh data lengkap yang dapat dianalisa secara sempurna. Kelemahan teknik imputasi telah menyebabkan adanya duplikat data, karena terdapat lebih dari satu data mempunyai nama yang sama, tetapi memiliki kelengkapan nilai yang berbeda. Metode Self Organizing Map Multiple Imputation (SOMMI) merupakan perbaikan dari metode Self Organizing Map Single Imputation (SOMSI) diusulkan untuk mengisi data yang hilang secara berulang kali (Multiple Imputation), dengan menggunakan beberapa bobot *centroid* untuk menangani data multivariate. SOMMI menggunakan teknik *ensemble* untuk menghasilkan beberapa bobot *centroid* pada kinerja clustering yang optimal. Hasil akhir imputasi dengan menggunakan bobot *centroid* akan menghasilkan data lengkap, yang akan dijadikan masukan pada klasifikasi Naïve Bayes (NB). Sehingga model yang menggabungkan SOMI dengan Naïve Bayes disebut dengan nama *Self Organizing Map Imputation Naïve Bayes* (SOMINB). Kelebihan pendekatan SOMINB dapat memanfaatkan beberapa nilai update bobot *centroid* untuk mengisi secara berulang setiap atribut yang terdapat nilai yang hilang. Pendekatan ini juga memiliki kelemahan yaitu adanya fitur dimensi tinggi yang menyebabkan masalah komputasi dan ketidakpastian nilai pengganti fitur pada pola data tertentu. Faktor fitur dimensi tinggi dapat diperbaiki dengan seleksi fitur dengan menggunakan Algoritma Genetika. Sehingga, penelitian ini memperkenalkan pendekatan baru yang menggabungkan prosedur preprosesing SOMI untuk mengatasi permasalahan data yang hilang dan optimasi *Naïve Bayes Classifier* (NBC) dengan Algorithm Genetika untuk seleksi fitur. Permasalahan lainnya, pemilihan nilai bobot optimal dan kesesuaian node jaringan menjadi masalah besar pada metode SOM. Algoritma Genetika juga digunakan untuk optimasi bobot pada clustering SOM, yaitu dengan menentukan bobot optimal di setiap cluster data. Bobot pada setiap cluster dipengaruhi oleh perubahan node output, selama proses pengelompokan pola input. Hasil penelitian dengan menggunakan model SOMINB menunjukkan nilai akurasi lebih tinggi senilai 90,00% dibandingkan dengan klasifikasi NB digabung dengan teknik imputasi lainnya. Pada tahap pengujian penggabungan preprosesing dengan Self Organizing Map Imputation, dan optimalisasi pemilihan fitur dengan Algoritma Genetika pada klasifikasi Naïve Bayes (SOMIGANB) telah menghasilkan akurasi 93,75. Model SOMIGANB lebih unggul dibandingkan penerapan pada klasifikasi data campuran kategori dan numerik lainnya. Hasil pemilihan bobot SOM pada missing value dengan menggunakan GA terbukti efisien dalam menemukan nilai bobot *centroid* untuk imputasi data multivariate.

Keywords: *Missing Data, NBI, Imputasi, Preprocessing, SOMI, SOMMI, Optimasi Bobot, Algoritma Genetika, Seleksi Fitur, Fitness Function*

ABSTRACT

The high-dimensional dataset has a common problem that is missing values caused by human error. The error occurred because the observation of the data was not going well, the measurement was incomplete, and the problem of respondents refused to answer certain questions in the survey. Missing value considered problematic, because the value has great influence on the making of decisions. Imputation techniques used to produce substitute missing values, in order to obtain complete data that can be analyzed completely. Weaknesses in imputation techniques have led to duplicate data, because there is more than one data that has the same name, but has a different set of values. Method of Self Organizing Map of Multiple imputation (SOMMI) is a refinement of the method Self Organizing Map Single imputation (SOMSI) proposed to fill in missing data repeatedly (Multiple imputation), using some weights centroid to handle multivariate data. SOMMI uses ensemble techniques to produce several centroid weights at optimal clustering performance. The final result of imputation using centroid weights will produce complete data, which will be input into the Naïve Bayes classification (NB). So the model that combines SOMI with Naïve Bayes is called the Self Organizing Map Imputation Naïve Bayes (SOMINB). The strength of the SOMINB approach is that it can utilize several centroid weight update values to repeatedly fill in each attribute that contains a missing value. This approach also has weaknesses, namely the presence of high-dimensional features that cause computational problems and the uncertainty of the replacement value of features in certain data patterns. High dimensional feature factors can be improved by feature selection using Genetic Algorithms. Since, this research introduces a new approach that combines SOMI preprocessing procedures to overcome the problem of missing data and optimization of Naïve Bayes Classifier (NBC) with Genetic Algorithms for feature selection. Another problem, the selection of the optimal weight value and suitability of the network node becomes a major problem in the SOM. Genetic algorithms are also used to optimize the weights in SOM clustering, by determining the optimal weights in each cluster data. The weight of each cluster is affected by changes in the output node, during the process of grouping input patterns. The results of the study using the SOMINB model show a higher accuracy value of 90.00% compared to the NB classification combined with other imputation techniques. In the testing stage, the merging of preprocessing with Self Organizing Map Imputation, and optimizing feature selection with Genetic Algorithms in the Naïve Bayes classification (SOMIGANB) has resulted in an accuracy of 93.75. The SOMIGANB model is superior to the application in the classification of mixed data and other numerical data. The results of the selection of SOM weights at the missing value using GA proved efficient in finding centroid weight values for multivariate imputation.

Keywords: *Missing Data, NBI, Imputation, Preprocessing, SOMI, SOMMI, Weight Optimization, Genetic Algorithms, Feature Selection, Fitness Function*