RINGKASAN

M. Aldiki Febriantono, Jurusan Teknik Elektro, Fakultas Teknik Universitas Brawijaya, Desember 2019, Klasifikasi Data *Imbalanced Multiclass* Menggunakan *Cost Sensitive Decision Tree C5.0* , Dosen Pembimbing 1: Dr. Ir. Sholeh Hadi Pramono, MS. Dosen Pembimbing 2: Rahmadwati, ST., MT., Ph.D.

Data mining merupakan proses penggalian data atau pencarian pola dengan tujuan mendapatkan informasi sebagai pengetahuan untuk mengambil keputusan secara cepat, tepat dan akurat di waktu yang akan datang. Pada proses data mining pendistribusian data yang tidak sama (*imbalanced data*) menjadi sebuah permasalahan yang penting karena *machine learning* lebih fokus pada *class* yang dominan (*mayority*) dibandingkan dengan *class* yang sedikit (*minority*). Padahal *minority class* dapat memiliki pengaruh yang jauh lebih besar jika terjadi salah klasifikasi (*misclassification*).

Dalam proses data mining, pola data dapat dipelajari sebagai dasar pengambilan keputusan. Salah satu cara untuk mencari pola data dengan menggunakan teknik klasifikasi. Klasifikasi memiliki dua jenis *class* prediksi yaitu *binary class* memiliki dua prediksi *class* dan *multiclass* memiliki prediksi *class* lebih dari dua. Metode klasifikasi yang sering digunakan untuk menyelesaikan masalah data *imbalanced* adalah *decision tree*. Konsep *decision tree* adalah merubah data berupa tabel menjadi model pohon kemudian menghasilkan aturan keputusan (*rule*).

Pada proses *decision tree* pemilihan atribut yang relevan memiliki pengaruh yang besar untuk mendapatkan performa yang baik. Hasil dari proses klasifikasi perlu adanya evaluasi untuk meminimalkan kesalahan klasifikasi. *Cost sensitive* merupakan metode yang mengansumsikan kesalahan klasifikasi sebagai *cost*. Metode tersebut bekerja dengan meminimalkan *cost* dari *classifier.*

Pada penelitian ini, tahap awal *dataset* dilakukan seleksi atribut menggunakan *particle swarm optimization*. Selanjutnya *decision tree* C5.0 digunakan untuk mencari pola pada *dataset* kemudian dilakukan pengujian. Hasil dari pengujian kemudian dilakukan evaluasi untuk mencari pola dengan nilai *cost* terkecil. Metode penelitian ini disebut dengan *cost sensitive decision tree* C5.0. Metode tersebut dilakukan pengujian dengan membandingkan decision tree C5.0 dengan ID3 dan C4.5. Selain itu *cost sensitive decision tree* C5.0 juga dibandingakan dengan *cost sensitive naïve bayes.*

Hasil pengujian klasifikasi *dataset* menggunakan *cost sensitive decision tree* ID3 tidak mampu meningkatkan nilai *accuracy* pada semua *dataset*. Hasil pengujian klasifikasi *dataset* menggunakan *cost sensitive decision tree* C4.5 mampu meningkatkan nilai *accuracy* pada tiga *dataset* antara lain *vehicle* dan *wine* berturut-turut meningkat sebesar 76.86% dan 97.62%. Sedangkan klasifikasi *dataset* menggunakan *cost sensitive decision tree* C5.0 mampu meningkatkan nilai *accuracy* pada *dataset glass* dan *thyroid*, berturut-turut sebesar 75.27% dan 95.81%. Hasil pengujian, *cost sensitive decision tree* C5.0 memiliki nilai *accuracy* yang lebih baik dari pada menggunakan metode *cost sensitive naïve bayes* pada *dataset glass, lympografi, vehicle* dan *wine* berturut-turut 76.17%, 83.33%, 75.27% dan 95.83%. Sedangkan dengan menggunakan metode *cost sensitive naïve bayes* memiliki nilai *accuracy* yang lebih baik dari pada *cost sensitive decision tree C5.0* pada *dataset thyroid* sebesar 97.67%.

Kata Kunci: *Data Mining, Cost sensitive, Decision Tree, Multiclass, Naïve Bayes.*

SUMMARY

M. Aldiki Febriantono, Department of Electrical Engineering, Faculty of Engineering, Universitas of Brawijaya, December 2019, Classification of Imbalanced Multiclass Data Using Cost Sensitive Decision Tree C5.0, Academic Supervisor 1: Dr. Ir. Sholeh Hadi Pramono, MS. Academic Supervisor 2: Rahmadwati, ST., MT., Ph.D.

Data mining is the process of extracting data or looking for patterns with the aim of getting information as knowledge for making decisions quickly, precisely and accurately in the future. In the process of data mining, the imbalanced data becomes an important problem because machine learning is more focused on the mayority class than the minority class. Even though minority classes can have a greatly influence if it occurs misclassification.

In the process of data mining, data patterns can be studied as a basis for making decision. Classification technique is used to make data patterns. Classification has two types of prediction classes, namely binary class and multiclass. The classification method that is often used to solve imbalanced data problems is the decision tree. Concept of decision tree change data into tree models and produce decision rules. The selection of relevant attributes has a big influence to get good performance. The results of the classification process need an evaluation to minimize misclassification. Cost sensitive is a method that assumes misclassification as cost. The method works by minimizing the cost of the classifier.

In this study, the attribute is selected using particle swarm optimization. Decision tree C5.0 is used looking for patterns in the dataset. The results of the testing are then evaluated to obtain patterns with the smallest cost value. This research method is called cost sensitive decision tree C5.0. The method is compared by the decision tree ID3 and C4.5. In addition, the cost sensitive decision tree C5.0 is also compared by cost sensitive naïve Bayes.

The results of testing the dataset classification using cost sensitive decision tree ID3 is not able to increase the accuracy value on all datasets. The results of testing the dataset classification using the cost sensitive decision tree C4.5 can increase the accuracy value of two datasets, including vehicle and wine, respectively by 76.86% and 97.62%. Whereas cost sensitive decision tree C5.0 can increase the accuracy value of the glass and thyroid, respectively by 75.27% and 95.81%. The test results, the cost sensitive decision tree C5.0 has better accuracy values than cost sensitive naïve bayes method in glass, lympography, vehicle and wine datasets respectively by 76.17%, 83.33%, 75.27% and 95.83%. Whereas cost sensitive naïve bayes method has a better accuracy value than cost sensitive decision tree C5.0 in the thyroid dataset of 97.67%.

Keywords: Data Mining, Cost sensitive, Decision Tree, Multiclass, Naïve Bayes.