ABSTRACT

High-dimensional data has large data complexity that leads more difficult, less effective and less efficient in information gathering processes. It also leads to overfitting that gives poorly classified results. It is possible that there is an unnecessary dimension in the whole dimensions. Therefore, it takes certain techniques to reduce the dimensions, so that the classification model becomes more easily be understood. Feature selection is one way to overcome the problem. Feature selection discards features that are irrelevant to the class to improve classifier performance.

In this study, a new method was developed, that was a hybrid method for handling high dimension data, especially in multiclass data. The hybrid method is a combination of the chi square filter method and the Genetic Algorithm wrapper method with the Multiclass Support Vector Machine (SVM) as the evaluator. Chi square was chosen for it was reliable when faced with multiclass problems, while Genetic Algorithm (GA) was considered capable of working well in conjunction with Multiclass SVM. This study used ten datasets taken from UCI repository.

In the first process, Chi Square reduces features that have a low degree of relevance to the class. Furthermore, the subset of features become a bait against GA and are selected using SVM multiclass as an evaluator. The feature subset with the highest accuracy is the optimal one. Four classifiers were used as the performance evaluators : Naïve Bayes, k-Nearest Neighbors, Simple CART, and Random Forest Tree. The results of the Chi-GA method show a significant increase of accuracy across most datasets. This proves that the Chi-GA method provides promising results on multiclass data handling.

Keywords – Chi Square, High dimensional data, Genetic Algorithm, Multiclass Support Vector Machine

INTISARI

Data dimensi tinggi mempunyai kompleksitas data yang besar sehingga berakibat pada proses penggalian informasi yang lebih sulit, kurang efektif, dan kurang efisien. Data dimensi tinggi juga menyebabkan terjadinya *overfitting* sehingga hasil klasifikasi menjadi jelek. Tidak menutup kemungkinan bahwa dari keseluruhan dimensi, ada dimensi yang tidak diperlukan. Oleh karena itu, diperlukan teknik tertentu untuk mereduksi dimensi agar model klasifikasi menjadi lebih mudah dipahami. Salah satu cara yang dapat dilakukan adalah dengan seleksi fitur. Seleksi fitur membuang fitur-fitur yang tidak relevan terhadap kelas sehingga mampu meningkatkan kinerja *classifier*.

Pada penelitian ini akan dikembangkan suatu metode baru, yaitu metode hibrid untuk penanganan data dimensi tinggi, khususnya pada data *multiclass*. Metode hibrid yang digunakan merupakan kombinasi dari metode *filter Chi Square* dan metode *wrapper* Algoritme Genetika (GA) dengan *Multiclass Support Vector Machine* (SVM) sebagai evaluator. *Chi Square* dipilih karena andal untuk kasus data *multiclass*, sedangkan GA dianggap mampu bekerja baik bersamaan dengan *Multiclass* SVM. Penelitian ini menggunakan sepuluh *dataset* yang diambil dari UCI *repository*.

Pada proses pertama, *Chi Square* mereduksi fitur yang mempunyai tingkat relevansi rendah terhadap kelas. Selanjutnya, hasil *subset* fitur menjadi umpan terhadap GA dan dievaluasi menggunakan *Multiclass* SVM. *Subset* fitur dengan akurasi tertinggi merupakan *subset* fitur optimal. Evaluasi kinerja menggunakan empat *classifier*, yaitu *Naïve Bayes*, *k-Nearest Neighbour*, *Simple CART*, dan *Random Forest Tree*. Hasil metode hibrid, Chi-GA menunjukkan peningkatan akurasi yang signifikan hampir di seluruh *dataset*. Hal ini membuktikan bahwa metode Chi-GA memberikan hasil yang menjanjikan pada penanganan *multiclass* data.

Kata kunci – Algoritme Genetika*, Chi Square*, Data dimensi tinggi, *Multiclass Support Vector Machine*