INTISARI

SUPPORT VECTOR MACHINE UNTUK *MULTICLASS IMBALANCED* PADA DATA DIMENSI TINGGI

Oleh

TUSRIANA RAHMATIKA

17/418731/PPA/05515

Perkembangan teknologi menghasilkan data dari berbagai bidang. Dalam kasus nyata, data dimensi tinggi dapat terjadi dimana $p > n$, dengan p adalah variabel atau fitur dan n adalah banyaknya observasi. Secara teori, bertambahnya jumlah fitur memberikan hasil klasifikasi yang akurat. Namun dalam praktiknya dengan jumlah observasi yang terbatas dan jumlah fitur yang sangat banyak mengakibatkan proses *learning* menjadi lambat dan model yang terbentuk menjadi *overfitting*. *Support Vector Machine* (SVM) merupakan salah satu metode *machine learning* yang memiliki hasil yang baik dalam hal klasifikasi dan prediksi. Prinsip dari metode SVM adalah melatih sekumpulan data dengan suatu algoritma untuk menghasilkan model klasifikasi yang dapat membantu memprediksi kategori dari data baru. Namun, kondisi *imbalanced data* mengakibatkan hasil klasifikasi lebih cenderung ke kelas mayoritas. Tesis ini membahas SVM kasus *multiclass* dan untuk meningkatkan performansi SVM dilakukan tahap *preprocessing* terdiri dari seleksi fitur dan penanganan *imbalanced* data. Seleksi fitur dilakukan menggunakan *Fast Correlation Based Filter* (FCBF) serta metode SMOTE, Tomek links dan *combine* (SMOTE+Tomek links) untuk mengatasi kondisi *imbalanced data*. Pada kasus *multiclass*, metode yang digunakan adalah *One Against One* (OAO). Data *microarray* leukemia terdiri dari kelas ALL-B, ALL-T dan AML dengan *imbalanced ratio* sebesar 4,25 digunakan untuk membandingkan performansi metode SVM *multiclass imbalanced*. Metode tersebut terdiri dari SMOTE SVM OAO, Tomek links SVM OAO dan *combine* SVM OAO. Evaluasi performansi menggunakan *accuracy*, F-*measure*, G-*mean*, and waktu. Berdasarkan *stratified* 3 *fold cross validation,* metode terbaik yang diperoleh adalah SMOTE SVM OAO.

Kata Kunci : *Support Vector Machine*, *Multiclass*, *Imbalanced*, Data Dimensi Tinggi

ABSTRACT

SUPPORT VECTOR MACHINE FOR MULTICLASS IMBALANCED IN HIGH DIMENSIONAL DATA

By

TUSRIANA RAHMATIKA

17/418731/PPA/05515

Technological developments produce data from various fields. In real cases, high dimensional data can occur where $p > n$, where p is the variable or feature and n is the number of observations. In theory, increasing the number of features gives an accurate classification result. But in practice with a limited number of observations and a large number of features resulting in a slow learning process and the model formed becomes overfitting. Support Vector Machine (SVM) is a machine learning method that has good results in terms of classification and prediction. The principle of the SVM method is to train a set of data with an algorithm to produce a classification model that can help predict categories from new data. However, the imbalanced data condition results in the classification results more likely to be in the majority class. This thesis discusses the SVM multiclass case and to improve the performance of the SVM the preprocessing stage consists of feature selection and imbalanced data handling. Feature selection is performed using Fast Correlation Based Filter (FCBF) as well as SMOTE, Tomek links and combine (SMOTE + Tomek links) methods to overcome imbalanced data conditions. In the case of multiclass, the method used is One Against One (OAO). Leukemia microarray data consisting of ALL-B, ALL-T and AML classes with a imbalanced ratio of 4.25 was used to compare the performance of the multiclass imbalanced SVM method. The method consists of SMOTE SVM OAO, Tomek links SVM OAO and combine SVM OAO. Performance evaluation uses accuracy, F-measure, G-mean, and time. Based stratified 3 fold cross validation, the best method obtained is the SMOTE SVM OAO.

Keywords: Support Vector Machine, Multiclass, Imbalanced, High Dimensional Data