INTISARI

IMPLEMENTASI *SUPPORT VECTOR MACHINES* DALAM KONSTRUKSI *OBLIQUE RANDOM FOREST* UNTUK KLASIFIKASI PADA DATA BERDIMENSI TINGGI

oleh
Irfan Wahyu Febrianto
16/398656/PA/17618

Random forest (RF) telah menjadi salah satu metode klasifikasi gabungan yang banyak menjadi perhatian oleh para peneliti *machine learning*. Prinsip dari RF menghasilkan suatu metode klasifikasi yang *robust* terhadap data *noise* serta tidak *overfit*. Fungsi pemisah pada tiap simpul RF menggunakan pemisah ortogonal yang tegak lurus terhadap sumbu atribut, yang jika diterapkan pada data dimensi tinggi dimana sangat mungkin terjadi banyaknya dependensi antar atribut akan menghasilkan performa klasifikasi yang kurang baik. Pada data dimensi tinggi sangat mungkin ada begitu banyak kombinasi antar atribut dan sayangnya RF tidak bisa mengeksploitasi situasi ini secara efisien. Pada skripsi ini, diajukan suatu metode *oblique random forest* (ORF) yang memiliki pemisah miring pada tiap simpulnya dimana digunakan kombinasi dari berbagai atribut secara langsung. Pemisah miring ini diharapkan dapat lebih mampu beradaptasi pada data dimensi tinggi. Metode yang digunakan untuk menemukan pemisah miring yang optimal dalam skripsi ini adalah *support vector machines* (SVM). Pada skripsi ini dilakukan analisis terhadap 4 data *microarray* kanker berdimensi tinggi dengan metode ORF serta dengan metode individunya yakni RF dan SVM. Dengan perbandingan nilai akurasi, presisi, sensitifitas, spesivisitas, dan skor F1, dihasilkan kesimpulan bahwa secara umum ORF memiliki performa yang lebih baik dibanding RF dan SVM.

Kata kunci : Data Dimensi Tinggi, *Random Forest*, *Support Vector Machines*, *Oblique Random Forest*.

ABSTRACT

IMPLEMENTATION OF SUPPORT VECTOR MACHINES ON CONSTRUCTION OF OBLIQUE RANDOM FOREST FOR CLASSIFICATION IN HIGH DIMENSIONAL DATA

by

Irfan Wahyu Febrianto

16/398656/PA/17618

Random forest (RF) has become one of ensemble classification method that has a great concern to machine learning researchers. The principle of RF produces a classification method that is robust against noise and does not overfit. The split function a each node of RF uses an orthogonal split that is perpendicular to the attribute axis. That kind of split, does not have high performance if applied to data with very high dimensions where it is possible that many dependencies between attributes. In high dimensional data it is possible that there are so many combinations between attributes and unfortunately RF can not effectively exploit this situation. This thesis propose an oblique random forest (ORF) method which uses oblique split at each node where a combination of various attributes is used directly. This oblique split is expected to be better adapted to high dimensional data. This thesis uses support vector machines (SVM) to find the optimal oblique split. This thesis analyze 4 microarray high dimensional cancer data using ORF method and the individual methods which are RF and SVM. By comparing the accuracy, precision, sensitivity, spesificity, and F1 score, the conclusion is that, in general, ORF has better performance than RF and SVM.

Keywords : High Dimensional Data, Random Forest, Support Vector Machines, Oblique Random Forest.