monsiti/iStock

# Siasat Ensemble untuk Memperoleh Model Prediktif Super

Pendekatan *ensemble* yang menggabungkan beberapa model prediktif bisa meningkatkan akurasi. Namun, korelasi antarmodel yang tinggi akan berpengaruh negatif sehingga perlu disiasati.

PENULIS

**BAGUS SARTONO**
Dosen di Departemen Statistika
Institut Pertanian Bogor.

**BERBAGAI** ide pendekatan *ensemble* dalam pemodelan prediktif telah banyak dikembangkan oleh para pakar dan praktisi *data science* dengan satu tujuan utama; memperoleh model dengan ketepatan prediksi yang tinggi. Teknik *ensemble* bekerja dengan menggabungkan hasil prediksi dari beberapa model (selanjutnya disebut sebagai model dasar) menjadi satu prediksi akhir. Proses penggabungannya bisa saja merupakan proses sangat sederhana seperti merata-ratakan (*averaging*) untuk variabel target numerik dan suara terbanyak (*majority vote*) untuk variabel target kelas. Ada juga proses penggabungan yang lebih rumit seperti penggunaan teknik optimasi tertentu sampai penggunaan algoritme pemodelan dengan menggunakan hasil prediksi model dasar sebagai prediktor.

## Ensemble Meningkatkan Ketepatan

Proses penggabungan hasil prediksi beberapa model dasar dalam banyak kasus mampu meningkatkan ketepatan prediksi seperti yang dilaporkan oleh banyak studi di berbagai jurnal ilmiah. Secara matematis, peningkatan akurasi dari pendekatan *ensemble* untuk kasus klasifikasi dapat diilustrasikan sebagai berikut.

Terdapat lima model dasar yang masing-masing memiliki akurasi sebesar 70%, dan nilai akurasi dalam hal ini diartikan sebagai berapa persen hasil prediksi kelas yang sama dengan kelas sesungguhnya. Lebih lanjut kelima model dasar itu diasumsikan bersifat saling bebas. Jika kelimanya digabungkan menggunakan pendekatan *majority vote*, maka prediksi dari *ensemble* akan menghasilkan akurasi yang tepat jika setidaknya

ada tiga model dasar yang tepat. Dengan menggunakan formula peluang binomial didapatkan bahwa akurasi dari *ensemble* kelima model dasar itu adalah:

$$\sum_{k=3}^{5}\binom{5}{k}\times 0{,}7^{k}\times 0{,}3^{5-k} = 83{,}6\%$$

Akurasi dari *ensemble* ini lebih tinggi dari akurasi model dasar. Ilustrasi ini setidaknya memberikan gambaran potensi tercapainya akurasi yang lebih tinggi jika pendekatan *ensemble* diimplementasikan dalam pemodelan prediktif.

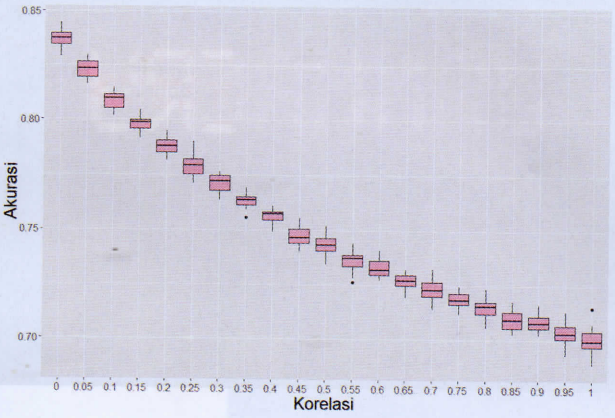
Peningkatan nilai akurasi bisa mencapai lebih tinggi lagi dibandingkan yang tertera di atas seandainya lebih banyak lagi model dasar yang digunakan. Misalnya saja untuk tujuh model dasar seperti di atas akurasi *ensemble* bisa mencapai 87%, untuk sembilan model dasar akurasinya menjadi 90%, dan untuk sebelas model dasar akurasinya bisa menjadi 92%.

Ekspektasi peningkatan akurasi inilah yang menjadi motivasi pengembang teknik *ensemble* untuk memperoleh model prediktif yang lebih baik dari model-model prediktif dasar alias model prediktif super.

## Kemungkinan Gagal

Perlu diingat bahwa pada saat kita menghitung ketepatan hasil prediksi dari pendekatan *ensemble* di atas, kita mengasumsikan antarmodel dasar bersifat saling bebas. Sederhananya, prediksi dari suatu model dasar tidak berkorelasi dengan prediksi yang dihasilkan oleh model dasar yang lain. Apa yang terjadi jika ternyata ada korelasi di antara model dasar?

Gambar 1 menampilkan hasil simulasi yang dilakukan oleh penulis terkait dengan efek




**Gambar 1.**
Grafik hasil simulasi terkait efek korelasi antarmodel dasar terhadap akurasi pendekatan *ensemble*.

korelasi antarmodel dasar terhadap akurasi pendekatan *ensemble*. Nilai rata-rata korelasi diatur sedemikian rupa mengambil nilai dari 0 sampai 1. Pengaturan dari *ensemble* yang disimulasikan serupa dengan rancangan yang diilustrasikan sebelumnya yaitu ada lima model dasar dengan akurasi masing-masing sebesar 70%.

Pada saat korelasi antara model dasar bernilai 0, tampak bahwa akurasi berkisar 83%-84% seperti yang dihasilkan perhitungan secara teoritis. Nilai akurasi akan cenderung terus menurun jika antarmodel dasar korelasinya meningkat. Ketika korelasi antarmodel dasar mendekati 1 atau berkorelasi sempurna, terlihat bahwa tidak ada gunanya melakukan proses *ensemble* yang ditandai dengan akurasi yang sama besar dengan akurasi model dasar yaitu 70%.

Simulasi ini setidaknya memberikan gambaran penting kepada kita bahwa pemilihan model dasar untuk digabungkan dalam proses *ensemble* penting untuk dikerjakan. Pasalnya, jika korelasi antarmodel dasar tinggi maka *ensemble* tidak akan memperbaiki tingkat akurasi pendugaan.

## Bagaimana Menyiasatinya?

Kenyataan bahwa korelasi antarmodel dasar berpengaruh negatif terhadap kinerja model *ensemble* sudah disadari oleh banyak pakar. Sebut saja bagaimana sejarah lahirnya pendekatan *random forest* dan *rotation forest* yang berupaya menutup kelemahan dari pendekatan *bagged tree*. Sebelumnya, *bagging* digadang-gadang menjadi pendekatan yang mampu memberikan perbaikan akurasi dari *classification tree*. Namun kemudian, disadari bahwa prediksi hasil banyak pohon klasifikasi yang dibentuk cenderung memiliki kemiripan yang tinggi, atau dengan kata lain korelasinya besar. Metode *random forest* dan *rotation forest* berupaya memperbaiki proses pembentukan pohon klasifikasi agar antarpohon memiliki korelasi yang lebih rendah. Caranya adalah dengan menggunakan sebagian prediktor saja secara acak pada saat membuat pohon dan melakukan rotasi agar diperoleh prediktor yang bersifat saling bebas.

Siasat lain dilakukan oleh Eric C. Polley dan Mark J. van der Laan dari University of California, Berkeley yang mengembangkan pendekatan *super learner*. Mereka mengusulkan untuk menggunakan berbagai model dasar, tetapi hanya beberapa model dasar saja yang digunakan dan diutamakan yang memiliki korelasi rendah. Berbeda dengan pendekatan *random forest*, *super learner* mengadopsi teknik *stacking* dan optimasi agar proses penggabungannya menghasilkan prediksi sebaik mungkin. IK