Creare calibrazioni stabili e robuste

Il tema di questo articolo è come fare calibrazioni NIR robuste e stabili. Dopo aver trattato alcuni argomenti generali relativi ai training set (set di addestramento) e allo sviluppo delle calibrazioni, descriverà specifici approcci alla protezione contro interferenze note.

Selezionare un set di campioni di addestramento

Il primo passo e il più importante quando si fanno calibrazioni NIR è la selezione dei campioni di addestramento. Qualsiasi algoritmo sia usato per costruire la regola predittiva, imparerà questa regola dai campioni di addestramento. Se la regola in seguito viene usata con campioni di un tipo non rappresentato nel training set, si possono avere prestazioni non soddisfacenti. Quindi il set di calibrazione ideale avrà non solo una buona presenza della proprietà di interesse, ma rappresenterà anche tutte le fonti importanti di variabilità che si troveranno nei campioni da prevedere in futuro. La difficoltà sta nell'identificare cos'è importante. Alcuni fattori, in particolare la dimensione delle particelle e il contenuto di umidità, sono importanti per molte applicazioni. Altri sono specifici di una determinata applicazione e possono richiedere uno studio maggiore. In generale, la migliore protezione contro l'eventualità di omettere accidentalmente una fonte importante di variabilità è quella di avere un training set più grande e variato possibile.

Scegliere e usare un metodo di calibrazione

Gli approcci standard di regressione dei componenti principali (principal components regression – PCR) e regressione ai minimi parziali quadrati, (partial least squares regression – PLSR) funzionano entrambi bene per la maggioranza delle applicazioni NIR di routine e dal mio punto di vista tra questi due c'è poco da scegliere. Quando l'intervallo della proprietà da prevedere è molto ampio, per esempio quando si tratta di composizioni in un intervallo da 0 a 40%, o quando il training set è molto eterogeneo, per esempio mangimi per animali con composizioni e forme fisiche molto diverse tra loro, le equazioni predittive lineari usate in questi approcci standard possono risultare carenti in flessibilità.

Allora l'uso di metodi locali, che si adattano alle equazioni lineari su intervalli limitati, o approcci non lineari, come le reti neurali artificiali (ANN) o le macchine a vettori di supporto (SVM) possono portare a migliori previsioni. Qualsiasi sia l'algoritmo usato, la cosa fondamentale per produrre una calibrazione stabile e robusta è evitare l'overfitting (eccessivo adattamento), usando o un test set o una qualche forma di cross-validation (validazione incrociata) per moderare il processo di adattamento. Più l'algoritmo è sofisticato, più diventa fondamentale evitare che si adatti troppo da vicino ai campioni su cui è provato diminuendo così la sua capacità di generalizzare. La procedura più sicura è usare un metodo più semplice possibile, e all'interno di tale metodo il modello più semplice possibile, evitando la tentazione di aggiungere tanta eccessiva complessità, per guadagnarci relativamente poco in termini di prestazioni.

Rendere la calibrazione robusta rispetto a fonti esterne di variabilità

Oltre alla variabilità naturale tra un campione e l'altro, anche altri fattori possono influenzare la calibrazione, come variazioni nella temperatura del campione e differenze tra gli strumenti usati per effettuare le misure degli spettri. Un approccio per rendere le calibrazioni robuste rispetto a questo tipo di variabilità esterna è semplicemente quello di includere la variabilità nel training set, per esempio, includere

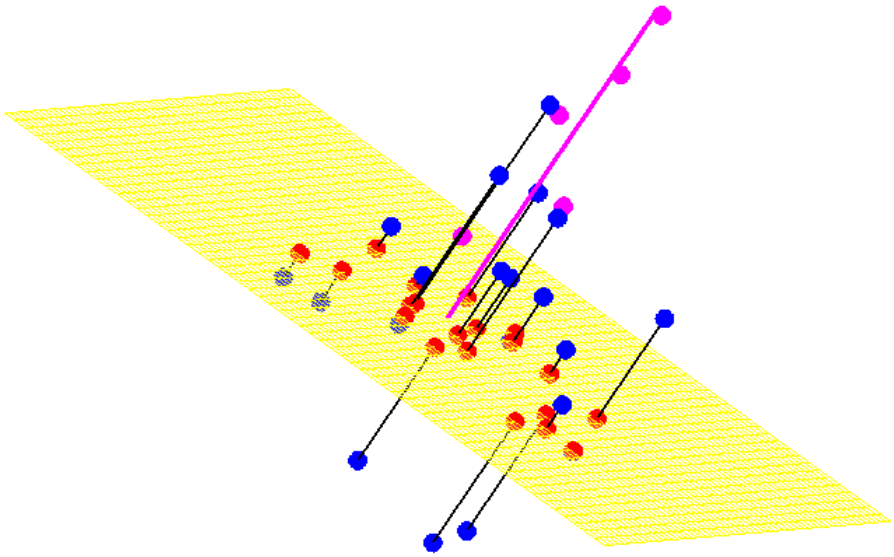

Fig. 1. Rimuovere una dimensione di rumore

campioni a temperature diverse o campioni misurati su strumenti diversi. Tuttavia quando siamo in grado di intervenire deliberatamente sui fattori di interferenza e quindi determinarne gli effetti in via sperimentale, diventano possibili altri approcci.

Supponiamo che il fattore di interferenza sia la temperatura. Allora l'esperimento necessario consiste nel misurare gli spettri di un numero ridotto di campioni ciascuno ad un numero ridotto di temperature. Separatamente per ciascun campione, creiamo un set di spettri di differenza, sottraendo uno degli spettri o lo spettro medio per quel campione da tutti gli altri. Poi uniamo questi spettri di differenza sui campioni per avere un set di spettri che colgano la variabilità spettrale causata dai cambiamenti nella temperatura. A questo punto ci sono due possibilità.

L'approccio "repeatability file"[1] in pratica aggiunge questi spettri al set di addestramento con riferimento al valore zero. Questo dice all'algoritmo di calibrazione che la variabilità spettrale di questo tipo non deve modificare le predizioni, e avrà un effetto simile a quello dell'approccio più ovvio, di includere campioni di temperature diverse nel training set. Il principale vantaggio di procedere in questo modo è che è facile, in questo quadro, dare agli spettri di differenza un peso maggiore nella calibrazione.

Questo, e le informazioni precise fornite dall'esperimento, indicano che il lavoro può essere fatto con un numero relativamente ridotto di misurazioni extra.

Una proposta alternativa[2,3], più recente è quella di usare gli spettri di differenza per identificare le direzioni nello spazio spettrale in cui si trova la maggior parte della variabilità dovuta alla temperatura ed applicare un pretrattamento agli spettri del set di addestramento che rimuova queste direzioni. Questo è illustrato in modo geometrico in figura 1. Ciascun punto nella figura 1 rappresenta uno spettro misurato a tre lunghezze d'onda (non molte, ma ci permettono di tracciare un quadro) e tracciato su uno spazio tridimensionale dove un asse corrisponde all'assorbenza a ciascuna lunghezza d'onda. I punti blu sono i campioni di addestramento. I punti viola sono gli spettri di differenza dall'esperimento in cui il fattore di interferenza è stato variato.

La linea viola è la direzione del primo componente principale da un'Analisi dei Componenti Principali (PCA) di questi spettri di differenza. Quasi tutta la variabilità spettrale dovuta al fattore di interferenza sta in questa direzione. Questo componente principale è usato per definire il piano giallo, che è ortogonale (ad angoli retti) ad esso, e poi gli spettri del set di addestramento sono proiettati su questo piano a dare i punti rossi con i punti sul piano più vicini ai punti blu a rappresentare gli spettri.

Per fare la calibrazione, usiamo gli spettri proiettati, cioè i punti rossi, invece degli spettri grezzi, cioè i punti blu.

Dal momento che questi spettri non hanno variabilità nella direzione che è più interessata dal fattore di interferenza, la calibrazione risultante sarà insensibile al fattore. I calcoli necessari per fare tutto ciò sono relativamente semplici: una PCA dei diversi spettri, e una proiezione ortogonale degli spettri di addestramento, che è una semplice moltiplicazione di matrici.

Quello che piace in questo approccio è che probabilmente porterà a modelli più semplici e più interpretabili perché sottrae la variabilità dell'interferenza invece di aggiungerla. Se dovessimo includere i punti viola nel set di addestramento ma dar loro valore di riferimento uguale a zero, un metodo di calibrazione come il PLS che costruisce i fattori includerebbe inevitabilmente nel suo spazio fattoriale la variabilità di interferenza, cioè la direzione viola.

Questa direzione otterrebbe poco peso nell'equazione di predizione, considerati i valori di riferimento zero, ma sembra più logico escludere la direzione da subito, invece di considerarla, includerla come modello nello spazio fattoriale e poi cercare di eliminarla.

Certo, ci sono dei limiti in quello che si può ottenere con qualsiasi metodo. Se tutta la variabilità spettrale dovuta al parametro per cui stiamo cercando di individuare la calibrazione cade esattamente nelle stesse dimensioni della variabilità spettrale dovuta al fattore di interferenza, allora eliminare o ridurre il peso di queste dimensioni distruggerebbe o danneggerebbe seriamente la calibrazione.

Per fortuna, uno dei grandi punti di forza del NIR come strumento di misura è che le informazioni tipicamente sono ripetute in parti diverse dello spettro, in modo che la sovrapposizione completa di variabilità da due fonte distinte qualsiasi sia rara, e che si possano usare in molte situazioni approcci come quelli descritti per rendere più robuste le calibrazioni senza intaccarne l'accuratezza.

Riferimenti

1. Westerhaus, M.O., in: Biston R. and Bartiaux-Thill (Eds), Proc. 3rd International Conference on Near Infrared Spectroscopy, Agric. Res. Ctr. Publ, Gembloux, Belgium, 1991.

2. Roger, J.-M., Chauchard, F. and Bellon-Maurel, V., Chemom. Intell. Lab. Syst., 66, 2003, 191-204.

3. Andrew, A. and Fearn, T., Chemom. Intell. Lab. Syst., 72, 2004, 51-56.

■

del Prof. Tom Fearn, Dipartimento di Scienze Statistiche, University College London