Rezumat

Există multe modele de învățare supervizată pentru clasificare (de ex. cele bazate pe rețele neuronale) a căror ieșire este o estimare a probabilităților *a posteriori* de apartenență la clasele considerate. Aceste modele antrenate pe un anumit set de date învață și distribuția acestuia. Ca efect, performanța maximă este garantată atunci când condițiile setului de test corespund celor din faza de antrenare. Un caz în care această condiție nu este îndeplinită este cel în care distribuția etichetelor datelor de test diferă față de cea a etichetelor datelor de antrenare, situație în care performanța poate fi diminuată. Această problemă poartă denumirea de „schimbare a probabilităților *a priori*" (eng. *prior probability shift*). Există câteva abordări pentru a o remedia, două dintre ele fiind exemplificate în această lucrare.

În cazul prezentat anterior, distribuția este privită ca o informație globală - setul de date de testare este disponibil complet și se pot încerca diferite ajustări pentru a îmbunătăți performanța dacă există o diferență de distribuție între cele două seturi de date. Există însă cazuri în care nu se dispune de întreg setul de date. Un exemplu este un model online care face predicții pe un flux continuu de date. În această situație, distribuția datelor primite până la un moment dat reprezintă o informație locală, dar care poate fi folosită pentru a influența predicția curentă, urmărind maximizarea performanței. Faptul că datele sunt continue conduce însă la apariția unor probleme, cum ar fi necesitatea unei estimări continue a distribuției locale. Dacă în cazul static, distribuția datelor de test era unică, acum fluxul de date poate proveni de la utilizatori diferiți, în intervale de timp diferite (de ex., un model care rulează pe un server pentru predicția vârstei unei persoane pe baza unei fotografii poate avea utilizatori preponderent dintr-un anumit interval de vârstă, în funcție de momentul în zi). Astfel, estimarea distribuției locale pe baza datelor disponibile până la momentul curent trebuie să trateze și cazurile în care aceasta poate varia în timp.

În această lucrare sunt construite câteva scenarii care să ilustreze problema de mai sus și se încearcă integrarea unei soluții clasice de ajustare a modelului în cazul continuu.