TiMBL – Tilburg Memory-Based Learner

Juraj Duráni

7.12.2012

TiMBL je program vyvinutý na Univerzite v Tilburgu pre NLP aplikácie. Jeho základ je Memory Based Learning (MBL), čo je jednoduchá a robustná metóda pre tento druh úloh.

MBL je následník k-NN kalsifikátora. Ten sa osvedčil pre vzory s numerickými atribútmi. V mnohých prípadoch pri NLP sa ale jedná o diskrétne atribúty, a veľké množstvo vzorov. Pri MBL je učenie veľmi rýchle, ale následná klasifikácia môže byť pomalá. Preto je tu prirodzená snaha o zrýchlenie a optimalizáciu tohto procesu. Výsledkom toho je v TiMBL vytvorenie stromovej architektúry. Tá umožňuje presné nájdenie k-NN, spolu s možnosťou pohybu v strome rovnako ako pri decision-tree klasifikátore.

Pre určenie podobnosti dvoch vzorov TiMBL v súčastnosti (máj 2011, verzia 6) implementuje 8 rôznych metrík (dve sú v štádiu rozpracovania) spolu so 7 rôznymi metódami pre určenie vzdialenosti v príslušných metrikách.

Základná metrika používaná v TiMBL je *Overlap metric*. Tá určuje podobnosť dvoch vzorov podľa jednoduchého vzorca. Pri nenumerických hodnotách je to 0, 1 v závislosti od toho, či sa hodnoty rovnajú alebo nie. Pri numerických je rozdiel počítaný presnejšie. Nevýhody pri nenumerických hodnôtách odstraňuje *Levenstheinova metrika* a *Diceov koeficient*, ktoré počítajú podobnosť dvoch slov.

Nie každý atribút musí byť ale rovnako dôležitý pre určenie triedy. K tomu v TiMBL slúži *Information Gain (IG)*. Ten určí každému atribútu jeho váhu na základe entropie v situáciách s a bez atribútu. Algoritmus využívajúci túto metriku sa nazýva IB1-IG (resp. IGTree).

V TiMBL sú implementované tieto algoritmy:

- IGTree - Využíva *Information Gain*, avšak štrukúra dát je komprimovaná do stromovej reprezentácie pri zachovaní informácie. Informácia z IG je používaná pri rozhodovaní o poradí hrán v strome. Podobné vzory majú podobnú cestu stromom. Síce potrebuje viac času pri učení, ale klasifikácia je následne rýchlejšia.

- TRIBL(2) - TRIBL a TRIBL2 algoritmy sú hybridy. Je to spojenie IB1 (ktoré nájde správne riešenie, ale je pomalé) a IGTree (ktoré je rýchle, ale v prípade, že IG nemá dostatočnú generalizačnú schopnosť zlyháva). Na začiatku sa používa IGTree na niekoľko atribútov a od určitého bodu sa použije klasický IB1. Rozdiel medzi TRIBL a TRIBL2 je v tom, kedy sa prejde na IB1 (TRIBL – fixný bod, TRIBL2 – dynamicky).

- IB2 - Ukladá do pamäte iba dôležité vzory. Pokiaľ sa vzor nedá klasifikovať podľa aktuálne uložených vzorov, je uložený. Je však citlivý na šum.

TiMBL je prístupný pod GNU. Aktuálna verzia je 6.4. Viac informácí o algoritmoch, metrikách a ich implementácii je možné nájsť v TiMBL Reference Guid. TiMBL je možné použiť vo vlastnom C++ programe.

Zdroje

Daelemans, W., Zavrel, J., Van der Sloot, K., and Van den Bosch, A. (2010). TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide. ILK Research Group Technical Report Series no. 10-01.