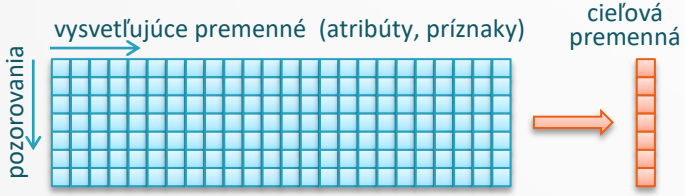
NÁVRH METÓDY VÝBERU PRÍZNAKOV NA BÁZE k-NN ALGORITMU

Peter Bugata | Peter Drotár | Fakulta elektrotechniky a informatiky | Technická univerzita v Košiciach

Problém

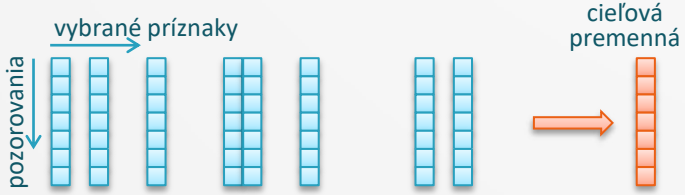
Pri objavovaní znalostí v dátach sa často využívajú algoritmy strojového učenia, ktoré na základe známych dát predpovedajú vlastnosti neznámych alebo nových údajov. Predpovedaná vlastnosť, tzv. cieľová premenná, závisí od ostatných - vysvetľujúcich premenných.



V súboroch dát sa obyčajne nachádza veľké množstvo atribútov, ktoré vytvárajú vysokorozmerný priestor. To zvyšuje požiadavky na výpočtové zdroje a predlžuje čas výpočtu. U niektorých algoritmov sa navyše vyskytuje jav známy ako prekliatie dimenzionality, ktorý výrazne znižuje presnosť predikcie.

Výber príznakov

Jedným z možných riešení uvedeného problému je predspracovanie súborov dát použitím metód výberu príznakov (*feature selection*), ktoré spomedzi všetkých vysvetľujúcich premenných vyberajú menšie podmnožiny premenných dôležitých pre presnosť predikcie.

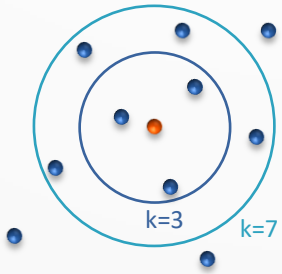


Obsahom diplomovej práce je návrh a implementácia novej metódy výberu príznakov na základe známeho algoritmu strojového učenia k-NN (*k* najbližších susedov).

Návrh metódy WkNN

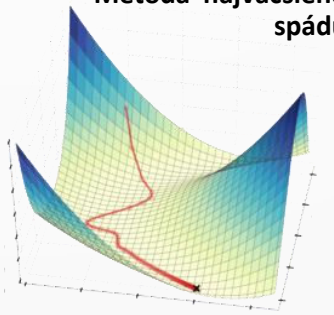
Pri návrhu metódy WkNN sa používa variant k-NN pre regresnú úlohu so špeciálnymi úpravami:

- použitím ohodnotenia susedov podľa vzdialenosti ,
- použitím váženia atribútov podľa ich dôležitosti pri určovaní hodnoty cieľovej premennej,
- využitím konceptu *leave-one-out* pre určenie chyby, kde sú namiesto *k* susedov uvažované všetky ostatné pozorovania.



Algoritmus k-NN

Metóda najväčšieho spádu



Minimalizáciou priemernej chyby predikcie metódou najväčšieho spádu (*gradient descent*) sú získané optimálne váhy, ktoré určujú dôležitosť jednotlivých atribútov pre presnosť predikcie. Atribúty s nulovými, resp. veľmi malými váhami možno vynechať.

Vlastnosti

Metóda WkNN bola navrhnutá ako všeobecný rámec so širokými možnosťami parametrizácie:

- použitím rôznych definícií vzdialenosti ,
- využitím rôznych funkcií chyby,
- použitím rôznych hodnotiacich funkcií vzdialenosti.

Pre získanie finálneho výberu premenných a ich váh je podporené použitie rekurzívnej eliminácie alebo regularizácie, ktorá zároveň zabraňuje tzv. pretrénovaniu.

Vyhodnotenie

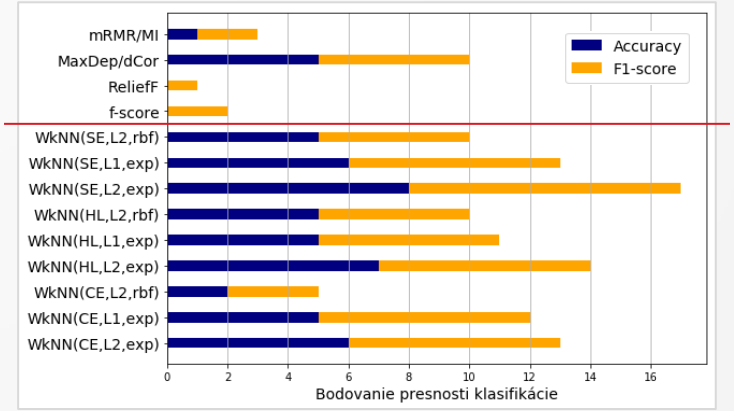
Navrhnutá metóda bola porovnávaná s inými známymi metódami výberu príznakov z hľadiska:

- identifikovania známych relevantných príznakov,
- vplyvu výberu príznakov na presnosť predikcie,
- stability výberu príznakov.

Na porovnávanie bolo použitých 5 syntetických súborov dát a 9 reálnych súborov dát získaných z DNA čipov.

Z hľadiska hľadania známych relevantných príznakov v syntetických dátach boli varianty navrhnutej metódy úspešnejšie ako všetky ďalšie porovnávané metódy.

Graf zobrazuje porovnanie metód na reálnych údajoch z hľadiska vplyvu na presnosť predikcie. Najúspešnejšie boli niektoré varianty navrhnutej metódy (pod čiarou).



Navrhnutá metóda je použiteľná pri regresii aj binárnej klasifikácii v kombinácii s rôznymi algoritmami strojového učenia. Dá sa aplikovať na súbory dát s vysokým počtom premenných a malým počtom pozorovaní, ktoré sa často vyskytujú v oblasti biomedicíny, ako aj na súbory s väčším počtom pozorovaní použitím stochastického variantu metódy najväčšieho spádu.