Použitie transformačnej regresnej techniky pre dolovanie v údajoch

Peter Krammer, Ladislav Hluchý

Ústav Informatiky
Slovenská Akadémia Vied
Dúbravská cesta 9, 845 07 Bratislava, Slovenská republika

{peter.krammer, ladislav.hluchy}@savba.sk

Abstrakt. Metódy dolovania a strojového učenia je možné aplikovať v mnohých doménach. Avšak viaceré spomedzi oblastí generujú len obmedzený objem dát, resp. získanie väčšieho objemu dát je drahé, časovo resp. technicky náročné. Aj v týchto oblastiach však vzniká potreba modelovania a predikcie, pričom nižší objem dát spôsobuje problematické zovšeobecnenie vlastností, čo sa prejaví nižšou presnosťou modelu. Článok pojednáva o dátovej transformácií, ktorá klasickú regresnú úlohu transformuje na úlohu s výrazne vyšším počtom záznamov s cieľom zvýšenia presnosti modelovania. Článok prináša modifikáciu dátovej transformácie ako aj jej otestovanie na reálnych dátových množinách. Pri tom porovnáva a hodnotí dosiahnutú výkonnosť natrénovaných modelov.

Typ príspevku: Výskumný príspevok

Kľúčové slová: dátová transformácia, regresia, modelovanie, dolovanie

1 Úvod

Súčasným trendom v mnohých častiach informatiky je rozhodne oblasť veľkých dát. Táto oblasť poskytuje výzvu pri riešení problémov efektívneho narábania so zdrojmi, škálovateľnosti, ako aj použitia vhodnej distribuovanej architektúry a podobne. Pri dolovaní v údajoch sa však často stretávame s opačným prípadom, kedy nemôžeme hovoriť o veľkých dátach; pri dostupnosti len niekoľko tisíc záznamov či dokonca menej. Takéto prípady nastávajú v doménach, kde meranie a zbieranie dát je časovo, alebo technicky náročné, resp. ekonomicky nákladné. Problémom sa tak skôr stáva reprezentatívnosť dátovej množiny a schopnosť generalizovania vzťahov modelom, čo sa prejaví znížením miery presnosti modelu. Aj napriek týmto problémom však vzniká potreba modelovania a predikcie veličín aj z týchto oblastí. V súčasnosti sa za účelom spresňovania modelov zvyčajne používajú metódy združeného učenia [1]. Tieto metódy často zlučujú viaceré rozdielne typy modelov, ktoré tak vzájomne kompenzujú svoje slabé stránky. Výsledný združený model, ktorý je zložený z čiastkových modelov

tak obvykle dosahuje vyššiu mieru presnosti predikcií, pretože predpovede sú výsledkom hlasovania čiastkových modelov. Druhým často používaným princípom v združenom učení[1] je viacnásobné trénovanie jedného typu modelu, pričom váhy jednotlivých záznamov sa menia v závislosti od úspešnosti predpovedí. Tento spôsob používa aj známa metóda AdaBoost. Viaceré spomedzi metód združeného učenia (Boosting, Bagging) boli pôvodne určené pre úlohu klasifikácie do tried, avšak neskôr boli navrhnuté aj modifikácie pre úlohu regresie [2], [6]. Celkovo však tieto metódy vychádzajú z princípu zlúčenia viacerých modelov, čo vedie k zložitej štruktúre vytvoreného modelu. V našom príspevku však používame len jeden model, ktorý je trénovaný na transformovaných dátach. Ďalšími výhodami tohto prístupu sú možnosť použitia dodatočného združeného učenia (pre ďalšie spresnenie), ako aj možnosť voľby typu použitého modelu.

1.1 Základný princíp transformácie

Základnou ideou transformačnej techniky je transformovať pôvodnú regresnú úlohu na ekvivalentnú s vyšším počtom záznamov a atribútov tak, aby tieto dáta boli svojou štruktúrou vhodnejšie pre proces strojového učenia. Za týmto účelom je použitá dátová transformácia, ktorá z pôvodnej dátovej množiny postupne vyberá všetky možné dvojice záznamov, pričom jedna dvojica vytvára jeden záznam transformovanej dátovej množiny. Z pôvodných N záznamov v dátovej množine získame N^2 - N záznamov v transformovanej množine. Počet vstupných atribútov sa zvýši dvojnásobne, keďže okrem príslušného atribútu bude zastúpená aj diferencia príslušného atribútu. Prezentovaná transformácia je vhodná v prípade úlohy regresie, výhradne pri spojitých numerických atribútoch, obzvlášť v prípade menších dátových množín.

Uvažujeme, že vstupné dáta, s homogénnou štruktúrou - v tvare tabuľky už boli predspracované a obsahujú vybrané relevantné vstupné atribúty. Po aplikovaní dátovej transformácie budú tieto údaje transformované do štruktúry obsahujúcej okrem pôvodných hodnôt aj ich diferencie. Cieľovou veličinou sa stane diferencia z pôvodných cieľových veličín. Pri predikcií je nutné opätovne realizovať dátovú transformáciu na predikovaný záznam, ktorý spárujeme so záznamami z trénovacej množiny. Získame tak väčší počet odhadov cieľovej veličiny, z ktorých následne určíme finálnu hodnotu cieľovej predikovanej veličiny. Podrobnejšie je dátová transformácia, jej základné aspekty, stratégie určenia cieľovej hodnoty ako aj proces predikcie popísané v publikácií [5].

Princíp ktorý umožňuje, aby táto technika dosahovala zlepšenie má niekoľko aspektov. V prvom rade, použitie rozdielov (diferencií) do určitej miery vyjadruje mieru vzdialenosti (distance) v jednotlivých atribútoch. V prípade ak 2 záznamy obsahujú výrazne podobné hodnoty vstupných atribútov, je veľmi pravdepodobné že aj ich cieľové atribúty budú mať podobné hodnoty (za predpokladu že vstupné atribúty sú relevantné). V prirodzených systémoch so spojitými veličinami sa veľmi často používajú prístupy, vyšetrujúce dopad zmeny vstupu na zmenu výstupu. Takéto prístupy využí-

[1] http://www.machine-learning.martinsewell.com/ensembles/ensemble-learning.pdf

vajúce diferencie boli použité aj pri modelovaní v rámci kauzálnej analýzy [3],[4]. Sledovanie nie len hodnôt, ale aj zmien hodnôt teda umožňuje spresnenie výsledného modelu. To súvisí aj s narábaním s hodnotami v procese trénovania. V procese trénovania sa modeluje závislosť medzi vstupom (vstupmi) a cieľovým atribútom. Avšak bežne používané spôsoby trénovania zvyčajne nezohľadňujú súčasne viaceré záznamy a už vôbec nie rozdiel medzi ich hodnotami. Je to však pochopiteľné, vzhľadom na fakt, že zohľadnenie takýchto rozdielov by bolo výrazne časovo náročné. Avšak, výnimku tvorí model k-najbližších susedov (ktorý do veľkej miery inšpiroval aj vznik tejto transformácie), ktorý síce model ako taký netrénuje, avšak zohľadňuje aj rozdiely hodnôt v atribútoch, z ktorých nakoniec počíta vzdialenosti záznamov.

V druhom rade sa jedná o štatistický fakt, keďže z väčšieho množstva nezávislých odhadov, dokážeme získať presnejšiu predpoveď cieľového atribútu. Väčší počet odhadov taktiež umožňuje použitie rozličných stratégií určenia finálnej predpovedanej hodnoty (aritmetický priemer, váhovaný priemer, odstránenie extrémov, výber najbližších záznamov, prípadne ich kombinácie).

V porovnaní s pôvodnou verziou prístupu [5] využívajúceho dátovú transformáciu, bolo vykonaných niekoľko zmien. V procese predikcie neboli použité všetky dostupné záznamy z trénovacej množiny (tak ako v pôvodnej verzii), ale len K záznamov s najnižšou euklidovou vzdialenosťou voči predikovanému záznamu. Hodnotu parametra K teda môžeme podľa potreby ladiť, pre dosiahnutie lepších výsledkov metódy; v našom prípade bola zvolená hodnota K = 10. Na vybraté záznamy bol aplikovaný natrénovaný regresný model, čím sme získali 2K odhadov cieľovej hodnoty. Ďalším rozdielom v porovnaní s predchádzajúcou verziou prístupu, je použitie váhovania pri priemerovaní získaných odhadov. Váhy jednotlivých záznamov boli určené na základe prevrátenej hodnoty vzdialenosti. Pre zabránenie deleniu nulou - v prípade ekvivalentných záznamov bola k vzdialenosti pripočítaná konštanta 0,01. Ďalším rozdielom, oproti pôvodnej verzii bolo použitie normalizácie vstupných atribútov dostupných údajov na interval 0 až 1. Dôvodom bolo zabránenie vplyvu rozdielnych rozsahov jednotlivých atribútov na určenie vzdialenosti dvojice záznamov.

Pre objektívnejšie zhodnotenie vhodnosti transformácie a výkonnosti modelov bola validácia vykonaná 10-násobne. Pri každom z 10 opakovaní, trénovacia množina pozostáva z náhodne vybratých záznamov spomedzi dostupných, pričom testovacia množina obsahovala zvyšné záznamy. Podmienky pri porovnávaní dosiahnutých výsledkov s použitím a bez použitia dátovej transformácie boli totožné (boli dokonca použité rovnaké seedy pre náhodný výber záznamov do trénovacej množiny). Taktiež z dôvodu vyššej objektívnosti boli použité 2 typy regresných modelov - neurónová sieť a stromový model M5P, ako aj viaceré dátové množiny. Pri experimentoch, bol okrem aspektu presnosti predikcie sledovaný aj časový aspekt predikcie, ako aj možnosť určenia intervalového odhadu cieľového atribútu pre konkrétny predikovaný záznam.

2 Dosiahnuté výsledky

Cieľom tohto článku je otestovať prezentovanú dátovú transformáciu s modifikovanou stratégiou určenia finálnej hodnoty na reálnych dátových množinách. Ako dátové množiny boli použité množiny Combined Cycle Power Plant Data Set[2] (označená ako PowerPlant) a Energy Efficiency[3]. Uvedené množiny obsahujú 9568 a 768 záznamov pri 4, resp. 8 číselných atribútoch. Pre obe dátové množiny boli realizované rovnaké experimenty. Z dátovej množiny bolo náhodne zvolených 200 záznamov, ktoré tvorili trénovaciu množinu.

V prvej fáze boli natrénované 2 typy regresných modelov (neurónová sieť a regresný strom) nad originálnymi dátami. Každé trénovanie bolo realizované 10 krát, s rozdielnymi hodnotami seedu, pre odlišné inicializačné nastavenie siete, ako aj odlišne zvolené záznamy v trénovacej množine. Na zvyšných záznamoch boli modeli validované, pričom z 10 opakovaní bol určený priemer. Celý tento proces bol opakovaný aj pre nižší počet záznamov - 195, 190, 185,... 80. V druhej fáze boli za rovnakých podmienok (pri použití rovnakých hodnôt seedu, ako aj rovnakých vybratých záznamov) vytvorené aj modeli z transformovaných dát.

Tabuľky Tab 1. a Tab 2. demonštrujú priemerné dosiahnuté presnosti modelov vyjadrené korelačným koeficientom (KK) a strednou kvadratickou chybou (SKCH). Priemerné hodnoty sú vypočítané vždy z 10 realizovaných opakovaní. Ako prvý model bola použitá viacvrstvová neurónová sieť perceptronov s 2 skrytými vrstvami, pričom aktivačnou funkciou bol sigmoid. Koeficient učenia bol zvolený 0.3 a maximálny počet epoch bol nastavený na 500. Druhým modelom bol regresný strom M5P [7] pri použití minimálneho počtu záznamov na list 4, s orezaním.

Tab 1. Porovnanie výkonnosti natrénovaných regresných modelov s použitím a bez použitia dátovej transformácie na dátovej množine PowerPlant.

		Model neurónová sieť		Stromový model M5P	
		140 záznamov	180 záznamov	140 záznamov	180 záznamov
s transform.	KK	0.9595	0.9646	0.9652	0.9629
	SKCH	4.8021	4.5278	4.4963	4.6251
bez transf.	KK	0.9656	0.9663	0.9636	0.9641
	SKCH	5.1683	5.0259	4.5938	4.5534

[2] ML Repository: http://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant

[3] Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Energy+efficiency

Tab 2. Porovnanie výkonnosti natrénovaných regresných modelov s použitím a bez použitia dátovej transformácie na dátovej množine Energy Efficiency.

		Model neurónová sieť		Stromový model M5P	
		140 záznamov	180 záznamov	140 záznamov	180 záznamov
s transform.	KK	0.9977	0.9993	0.9982	0.9993
	KCH	0.8642	0.6981	0.8308	0.7035
bez transf.	KK	0.9776	0.9862	0.9804	0.9856
	KCH	2.2451	1.9694	2.1177	1.8697

Trénovacie množiny pozostávali zo 140 resp. 180 náhodne vybratých záznamov, testovacie množiny obsahovali zvyšné nepoužité záznamy. Ako modeli boli použité viacvrstvové neurónové siete perceptronov, ako aj regresné stromy M5P. Pri trénovaní modelov bola použitá knižnica Weka. Pri validácií boli vyčíslené kritériá korelačný koeficient (KK) a stredná kvadratická chyba (SKCH). Výsledné výkonnosti uvedené v tabuľkách predstavujú priemer z 10 validácií.

Z dosiahnutých výsledkov v Tab. 2 si môžeme všimnúť pri použití transformácie výrazný nárast presnosti natrénovaných modelov v oboch sledovaných kritériách. V tabuľke sú uvedené iba prípady s počtom záznamov 140 a 180, avšak aj ostatné testované prípady s počtom záznamov 80 až 200 dosahujú výrazne podobné výsledky.
Pri testovaní transformácie na dátovej množine PowerPlant, ktorej výsledky sú uvedené v Tab. 1, spresnenie nie je až natoľko výrazné. Obzvlášť je to zrejmé v prípade použitia regresných stromov M5P. Pri použití neurónových sietí sa so zvyšujúcim počtom záznamov v trénovacej množine zvyšuje aj presnosť modelu. Celkovo však modeli natrénované za použitia dátovej transformácie dosahujú v priemere lepšiu výkonnosť, len v niekoľkých individuálnych prípadoch dosiahli o niečo horšiu výkonnosť.

Z časového aspektu, je trénovanie a predikcia za použitia prezentovanej dátovej transformácie výrazne časovo náročnejšia. Tento aspekt je však očakávaný, vzhľadom na potrebu viacnásobného aplikovania modelu ako aj dátovej transformácie na predikované dáta. Predikcia za použitia transformácie je približne stonásobne pomalšia, v závislosti od typu stratégie určenia finálnej predikovanej hodnoty a počtu odhadov.
Prezentovanú techniku je preto vhodné použiť v prípade, ak primárnym kritériom je vysoká presnosť modelu a prípadné vyššie časové nároky nie sú prekážkou.

3 Záver

Celkovo, prezentovaná transformačná technika vykazuje potenciál, spočívajúci v zvýšení presnosti regresných modelov. Ukázalo sa to na syntetických [5] ako aj reálnych dátach, pričom zlepšenie presnosti modelu bolo zrejmé z oboch sledovaných kritérií - korelačného koeficientu, ako aj strednej kvadratickej chyby. Z časového hľadiska, použitie tejto techniky značne zvyšuje časovú náročnosť (obzvlášť vo fáze predikcie), čo je však efekt, ktorý bol pri návrhu techniky očakávaný. Je preto vhodné zvážiť použitie tejto techniky v závislosti od požiadaviek na presnosť a rýchlosť predikcie modelu, ako aj počtu záznamov v trénovacej množine. Celkovo však prezentovaná transformačná

technika vykazuje viacero pozitívnych aspektov, medzi ktoré patria aj možnosť voľby typu modelu, možnosť realizácie intervalového odhadu cieľovej hodnoty, možnosť voľby stratégie určujúcej výpočet cieľovej hodnoty ako aj výrazné spresnenie modelov. Je zrejmé, že nie u všetkých reálnych dátových množín dôjde k takto výraznému zvýšeniu presnosti. Do budúcnosti tak zostáva potreba podrobnejšie otestovať techniku na ďalších dátových množinách. Zaujímavé by tiež bolo porovnanie presnosti modelov používajúcich prezentovanú techniku a metódu boostingu.

Poďakovanie: Táto publikácia vznikla vďaka podpore projektu VEGA 2/0167/16.

Literatúra

1. Dietterich Thomas G.: Ensemble Methods in Machine Learning, Oregon USA, 1998.
2. Elith Jane, Leathwick John: Boosted Regression Trees for ecological modeling, 2016.
3. Kvassay M., Hluchý L., Krammer P., Schneider B.: Causal analysis of the emergent behavior of a hybrid dynamical system. In Acta polytechnica Hungarica: journal of applied sciences at Budapest Tech Hungary, 2014, vol. 11, no. 4, p. 21-40. (0.471 - IF2013). ISSN 1785-8860.
4. Kvassay M., Krammer P. Hluchý L., Schneider B.: Causal Analysis of an Agent-Based model of Human Behaviour, Computing and Informatics, 2016, vol. 32. (in review)
5. Krammer P., Hluchý L.: Transformačná regresná technika pre dolovanie v údajoch. WIKT 2014: 9th Workshop on Intelligent and Knowledge Oriented Technologies, Bratislava, p. 45-50, ISBN 978-80-227-4267-2.
6. Schonlau Matthias R.: Boosted Regression (Boosting), The Stata Journal 5, Number 3, 2005, pp. 330 - 3654.
7. Wang Y., Witten I. H.: Induction of model trees for predicting continuous classes, In Poster papers of the 9th European Conference on Machine Learning, 1997.

Annotation:

Using Transformation Regression Technique for Data Mining

Data mining and machine learning methods can be used in many domains. However, several domains generate limited volume of data only, because getting larger data sets is difficult from time, economical, or technical aspects. But these domains also require a modelling and predicting; so the small data volume can cause problems in generalization and decreasing of model precision. Presented paper deals about data transformation, which original regression task replace with regression task, with higher count of records, with tendency to increase model precision. Paper demonstrate a new modification of data transformation which testing on real data sets. Reached performance comparison and evaluation are published in paper.