Anachronické atributy a dobývání znalostí

Lenka Nováková, Jiří Kléma, Olga Štěpánková

Katedra kybernetiky, FEL, ČVUT - České vysoké učení technické v Praze,
Technická 2, Praha 6
novakova, klema, step@labe.felk.cvut.cz

Abstrakt Při zpracování lékařských dat získaných dlouhodobým sledováním pacientů se často setkáváme s tím, že se rozsah dat pro jednotlivé pacienty výrazně liší. Data o sledovaném souboru jsou pak nejednotná, přičemž zařazení pacienta do cílové třídy může přímo souviset s počtem dostupných měření o tomto pacientovi. Ovšem u této hodnoty je nebezpečí, že jde o anachronický atribut a nelze se o ni opírat při konstrukci libovolného modelu sledovaných dat. Příspěvek navrhuje možný postup, jak předzpracovat výchozí soubor dat tak, aby byl tento zásadní problém odstraněn. Navržená metoda je ilustrována na případu dat ze studie STULONG.

Klíčová slova: Anachronické atributy, Předzpracování dat, Časová řada

1 Úvod

U prediktivních úloh dolování dat, které pracují s časovými řadami měření, existuje reálné nebezpečí, že se v datech mohou vyskytnout tzv. *anachronické atributy* [6], [4]. Takto jsou označovány atributy, které jsou sice obsaženy v trénovacích datech, ale určitě nemohou být k dispozici v době budoucí požadované předpovědi. Jako příklad uveďme predikci tržeb v obchodě: ta nepochybně souvisí s tím, kolik ten který den přijde do obchodu lidí. Pokud použijeme pro předpověď tržby jako jeden z atributů „celkový počet návštěvníků obchodu v daný den“, budou naše předpovědi jistě poměrně kvalitní. Problém spočívá ovšem v tom, že počet lidí, kteří během dne přijdou, předem (ráno) neznáme. Jedná se o typický anachronický atribut, který není vhodné použít pro vytváření modelu pro predikci denních tržeb. V tomto případě by asi bylo lépe pokusit se nejprve nalézt metodu pro predikci „celkového počtu návštěvníků obchodu v daný den“ třeba pomocí dne v týdnu, ročního období, počtu zákazníků v daném období minulý rok apod. Teprve hodnotu tohoto predikovaného atributu by bylo možné přímo použít v modelu pro predikci denních tržeb.

Je zřejmé, že nemá smysl vytvářet model za pomoci anachronických atributů. Ovšem mnohdy bývá velmi obtížné tyto atributy rozpoznat, neboť tato jejich vlastnost nebývá explicitně zmíněna nebo ji lze v rozsáhlém popisu vstupních dat lehce přehlédnout. Navíc anachronický atribut může být v souboru dat přítomen nejen přímo, ale může se do souboru vloudit i při běžném způsobu

předzpracování, jakým jsou například různé postupy agregace dat. Problém tohoto typu budeme nejprve dokumentovat na konkrétní úloze. V závěru se pokusíme navrhnout postup, který považujeme za vhodný pro zpracování některých typů časových řad a který se vyhýbá nebezpečí zanesení anachronických atributů.

1.1 Projekt STULONG

Problém odvozených anachronických atributů velmi plasticky vykresluje analýza dat projektu STULONG [8].

Studie (STULONG) byla realizována na II. interní klinice, 1. lékařské fakulty UK a Všeobecné fakultní nemocnice, U nemocnice 2, Praha 2 – pod vedením prof. MUDr. F.Boudíka, DrSc., MUDr. M.Tomečkové, CSc. a doc. MUDr. J.Bultase, CSc. Většina dat byla převedena do elektronické podoby v rámci evropského projektu Managing Uncertainity in Medicine programu Copernicus na pracovišti EuroMISE (Evropského centra medicínksé informatiky, statistiky a epidemiologie) Karlovy univerzity a Akademie věd (pod vedením prof. RNDr. J.Zvárová, DrSc.). Analýza dat vznikla za podpory grantu MŠMT ČR LN 00B 107.

Jedná se o data z rozsáhlé epidemiologické studie primární prevence aterosklerózy, nazvané Národní preventivní multifaktoriální studie srdečních infarktů a cévních mozkových příhod. Studie zahrnuje dvacetileté pozorování přibližně 1400 mužů středního věku. Cílem projektu je identifikovat rizikové faktory aterosklerózy.

Data jsou rozdělena do 4 tabulek. Pro nás budou nejdůležitější dvě z nich - Entry tabulka se vstupními daty o pacientech a Control tabulka obsahující série kontrolních vyšetření. Po sloučení tvoří tyto dvě tabulky časovou řadu vyšetření o jednotlivých pacientech. Důležité je, že data o jednotlivých pacientech nemají zcela jednotný charakter. Předmětem sledování při opakovaných vyšetřeních všech pacientů byly sice stejné atributy, ovšem počet kontrol jednotlivých pacientů se výrazně různí, viz. obrázek 3, nebo se stává, že hodnoty některých atributů při části kontrolních návštěv chybí, tj. celkový počet kontrol je u pacienta vyšší než počet měření dané veličiny. Za těchto okolností je nutné časovou řadu dat o kontrolních vyšetřeních jednotlivých pacientů předzpracovat tak, aby informace o jednotlivých pacientech měly uniformní strukturu. Teprve pak bude možno použít pro modelování nejčastějších klasických postupů, které vycházejí z metod atributového strojového učení.

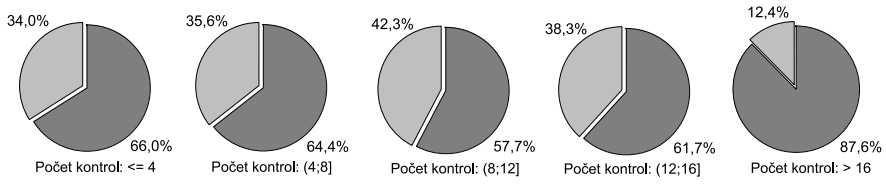
Vzhledem k tomu, že máme k dispozici různý počet kontrolních vyšetření pro různé pacienty, zdá se být přirozené nahradit řady měření odlišné délky jednou nebo více agregovanými hodnotami, např. průměrem nebo extrémními hodnotami. Podobné postupy při zpracování lékařských dat lze nalézt například v [3]. Za jeden typ rizikových faktorů se v dané úloze považují trendy vývoje sledovaných atributů. Trendy lze dobře charakterizovat pomocí parametrů regresní přímky proložené dostupnými časovými daty. Ovšem právě při výpočtu trendů může být do dat zanesena anachronická informace. Povšimněme si odpovídajících vzorců uvedených v odstavci 1.3, kde se opakovaně vyskytuje celkový počet dostupných měření. Celkový počet měření samozřejmě je anachronický atribut,

neboť u nového pacienta nemůžeme předpovídat, kolikrát ještě přijde. Ovšem regresní parametry vypočtené ze všech dostupných dat nemusí nutně být anachronické, neboť i pro nového pacienta je možné regresní parametry spočítat už po prvních 2 kontrolách. Je ale rozumné srovnávat regresní parametry vypočtené z trénovacích dat s regresními koeficienty nových pacientů vypočtenými z jejich dosavadních návštěv? K této otázce se vrátíme na konci odstavce 1.3.

1.2 Ověření závislosti mezi počtem kontrol a četností kardiovaskulárních onemocnění

Jednou z klíčových událostí studie je objevení kardiovaskulárního onemocnění (KVO) u daného pacienta. Vytvoření modelu, který odliší pacienty s KVO a bez KVO (nonKVO), je tedy důležitou úlohou v rámci projektu STULONG. Uvažujme binární atribut KVO jako cílovou veličinu vytvářeného modelu.

K ověření závislosti mezi počtem kontrol a KVO v průběhu studie se nabízí celá řada možných testů a vizualizačních nástrojů. My jsme využili nejprve kritéria plochy pod ROC křivkou (AUC - area under curve, ROC - receiver operating characteristic). AUC kvantifikuje schopnost separace dvou skupin klasifikačním modelem za různých podmínek. Uvažujme triviální model KVO založený pouze na počtu kontrol - AUC pak vyjadřuje pravděpodobnost správného zařazení pacienta do skupin KVO a nonKVO pro všechny možné prahové hodnoty počtu kontrol. Náhodný model vykazuje AUC 0.5, dokonalý model 1 a dokonalý inverzní model 0. My jsme AUC(počet kontrol, KVO)=0.38 odhadli pomocí neparametrické Wilcoxonovy statistiky [7], model pacienta správně zařadí s pravděpodobností 0.38 (95% interval spolehlivosti je [0.33,0.42]). Je tedy zřejmé, že četnost KVO významně klesá s rostoucím počtem kontrol (AUC<0.5). Podobně je možné zamítnout hypotézu o nezávislosti počtu kontrol a KVO pomocí χ^2 testu nezávislosti (na hladině významnosti 0.005). Vizualizace zmíněné závislosti pomocí kategorizovaných koláčových grafů je na obrázku 1. Graf naznačuje sníženou četnost onemocnění u skupiny pacientů s více než 16 kontrolami.



Obrázek 1. Rozložení KVO pro různé počty kontrol, tmavě KVO, světle nonKVO

Pozorovanou souvislost mezi KVO a počtem kontrol lze chápat jako důsledek metodiky použité při sledování pacientů. Sledování probíhalo téměř pravidelně každý rok a bylo ukončeno po 15 - 20 letech ukončením celé studie. U řady pacientů však bylo sledování ukončeno z nejrůznějších důvodů i dříve. Z našeho pohledu je podstatné, že pokud pacient onemocněl některým ze sledovaných kardiovaskulárních onemocnění, byl vyřazen z původní skupiny a dále sledován

na klinice v rámci sekundární prevence aterosklerotických onemocnění. V dalších letech s ním nebyl vyplňován kontrolní dotazník a z pohledu studie bylo jeho sledování ukončeno. Tento postup logicky může vést k tomu, že pacienti s mnoha kontrolami jsou s větší relativní četností pacienti zdraví a ti pacienti, kteří byli měřeni krátce, naopak častěji onemocněli některou sledovanou chorobou.

Je tedy zřejmé, že počet vyšetření v průběhu studie může být kauzálně ovlivněn případným onemocněním pacienta. Uvažujeme-li pacienty mimo studii, tedy objekty, na které bude model perspektivně aplikován, údaj o počtu vyšetření do možného onemocnění buď nemá žádný smysl nebo je přinejmenším neznámý. Počet měření v analyzovaných datech považujeme tedy za anachronický atribut, jehož zařazení do modelu je nutné se vyvarovat. Nejde přitom pouze o zařazení bezprostřední, ale i možné interakce s dalšími odvozenými atributy.

1.3 Použití agregovaných hodnot pro opakovaná měření

Při zpracování časových řad se často používá náhrada řady několika charakteristickými hodnotami jako je aritmetický průměr, směrodatná odchylka, případně náhrada regresní křivkou, ať již lineární nebo vyššího řádu. Jako příklad agregovaných hodnot uveďme výpočet parametrů regresní přímky $y = kx + q$:

$$k = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2},\ q = \frac{\sum_{i=1}^{n} y \sum_{i=1}^{n} x^2 - \sum_{i=1}^{n} x \sum_{i=1}^{n} xy}{n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2}$$

Z uvedených příkladů vztahů je vidět závislost, ať již přímou nebo nepřímou, těchto charakteristických hodnot na počtu měření n. Protože počet měření n je anachronický atribut, nelze vyloučit přenos anachronismu na takto nově vypočtené charakteristické hodnoty. Experimenty na STULONG datech prokázaly, že závislost pozorovaná v odstavci 1.2 se přenesla i na regresní koeficienty. Jinými slovy, bylo prokázáno, že hodnoty koeficientů závisí na počtu měření, ze kterých byly odvozeny. Jejich možná souvislost s objevením KVO může být pouze závislostí zprostředkovanou počtem měření n. Proto nelze použít agregované atributy vypočtené jako parametry regresní přímky pro budování prediktivního modelu KVO.

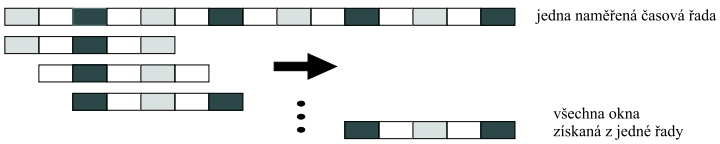
Proto se vstupní soubor dat musí transformovat tak, aby predikovaný výsledek nebyl na počtu měření závislý a teprve po takovéto úpravě můžeme již časové řady nahradit některou z uvedených nebo i dalších na počtu měření závislých charakteristik. Transformací, která tento problém řeší, je použití *časového okna*.

2 Časová okna

Metoda oken (windowing) je jednoduchým a často používaným postupem transformace časových dat, přehled různých metod pro dolování časových dat lze

nalézt např. v [1]. Konkrétní implementace se může lišit v závislosti na typu zpracovávané časové posloupnosti, popřípadě specifikaci úlohy. Může být chápána jako rozklad původní posloupnosti na větší či menší počet disjunktních oken nebo také jako metoda *klouzavého okna* (sliding window), při které se jednotlivé podposloupnosti překrývají. Typickým příkladem prvního postupu může být rozdělení časových dat na okna, která mohou být dále reprezentována lineární aproximací dat původních nebo na zcela symbolické úrovni. V obou případech získáme zjednodušenou reprezentaci vstupní sekvence využitelnou například pro efektivnější definici míry podobnosti při shlukování.

V případě naší úlohy nejde prioritně o vzájemnou podobnost časových řad odpovídajících různým pacientům, ale o souvislosti mezi vývojem jednotlivých rizikových faktorů v čase a případným onemocněním. V této situaci se nejvhodnější metodou jeví klouzavé okno pevné délky, které v každém časovém okamžiku vyjadřuje poslední vývoj sledovaného faktoru a aktuální zdraví pacienta. Obecně řečeno, metoda posuvného okna transformuje časovou řadu o n měřeních na novou sadu časových řad o konstantním počtu měření l. Pro generování těchto nových dat lze použít ty prvky původních dat, které obsahují alespoň l měření - ostatní prvky s méně měřeními je nutné vynechat. Po použití této úpravy přestává být predikovaný výsledek závislý na počtu měření, neboť v sadě nově upravených dat je počet měření konstantní. Postup transformace časové řady bez chybějících hodnot je znázorněn na obrázku 2.

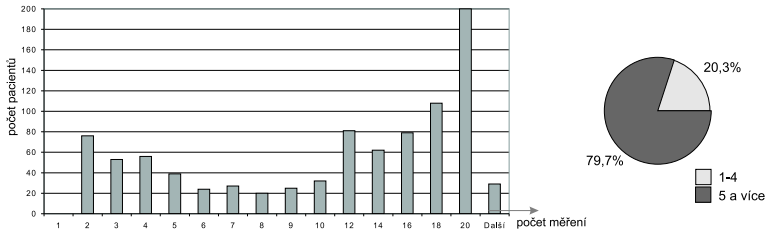


Obrázek 2. Transformace vstupní časové řady na okna

Metoda je parametrická, výsledek může být výrazně ovlivněn volbou délky okna l. Bohužel neexistuje univerzální předpis, jak tuto délku zvolit. Délka se volí s ohledem na řešenou úlohu, nejčastěji jako minimální časové období, kdy očekáváme v časové řadě změny, které umožní predikovat výsledek. Volba délky okna je velmi důležitá, protože na ní závisí prediktivní schopnost budoucího modelu, ta bude pro každou délku okna jiná.

Po stanovení velikosti okna provedeme vlastní transformaci. Místo každé časové řady získáme $n - l + 1$ nových řad, kde n je počet měření původní řady a l zvolená délka okna.

Protože se počet měření v našem případě pohybuje od 2 do přibližně 20 měření, je volba délky okna kompromisem mezi ztrátou dat a délkou sledovaného obdobíV. V řešené úloze jsme zvolili délku okna 5 měření, respektive 5 let (viz dělení v kapitole 3). Histogram a koláčový graf na obrázku 3 naznačují, že při této volbě velikosti okna ztrácíme data asi o jedné pětině pacientů. Z histogramu zároveň plyne, že alternativní volbou by mohla být i okna o délce 8 nebo 10 měření, ale všechna delší okna už zanedbávají více než polovinu pacientů.



Obrázek 3. Histogram počtu měření u dat STULONG

2.1 Vývoj klasifikace sledovaných objektů

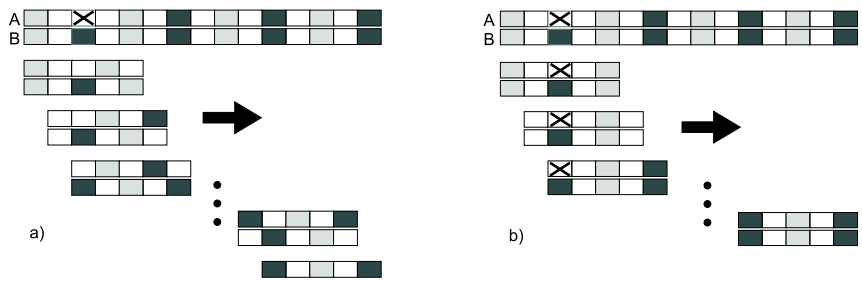
V odstavci 1.2 jsme veličinu KVO chápali jako nezávislou na čase. Se zavedením časových oken je nutné ji vztáhnout k aktuálnímu času. Největší objem informace zůstane zachován, pokud původně binární atribut KVO nahradíme celočíselnou hodnotou, která ve sledovaném případě odpovídá době, která uběhne mezi aktuálním okamžikem (koncem daného okna) a zjištěním KVO u daného pacienta. V případě, že KVO nebylo u pacienta vůbec diagnostikováno, zůstává tato hodnota nevyplněna. Rozhodnutí o způsobu náhrady chybějící hodnoty záleží na expertu z oblasti úlohy. V případě dat STULONG byla chybějící hodnota nahrazena hodnotou „zdráv“.

3 Metoda oken a chybějící hodnoty

Metoda oken pevné délky umožňuje použít pro modelování řadu odvozených atributů, např. atributy popsané v odstavci 1.3. Ovšem pokud pracujeme s časovými řadami s chybějícími hodnotami, musíme i v tomto bodě postupovat dostatečně obezřetně. Vyskytují-li se v datech chybějící hodnoty můžeme zachovat buď pevný počet měření - okno obsahující chybějící hodnoty doplníme následujícími hodnotami prostým posunutím řady, nebo pevnou časovou délku okna - chybějící hodnoty buď ignorujeme (agregát je založen na menším počtu hodnot) nebo raději navrhneme způsob jejich náhrady.

3.1 Náhrada chybějící hodnoty posunutím řady

Máme-li časovou řadu s chybějícími hodnotami, můžeme chybějící hodnotu vynechat a okno doplnit hodnotou následující. Tento postup má ale jedno velké úskalí, které vysvětlíme na příkladě. Budeme mít objekt, který je popsán dvěma časovými řadami, jedna bude reprezentovat veličinu (atribut) A a druhá veličinu B, v obrázku A je křížkem označena chybějící hodnota. Při transformaci nahradíme tuto chybějící hodnotu hodnotou následující. U atributu B žádná hodnota nechybí, proto k posunu nedochází. Výsledkem je, že pro atribut A máme o jedno klouzavé okno méně a jednotlivá okna za náhradou posunutím si neodpovídají. Tento postup není z uvedeného důvodu použitelný, stejnolehlé agregované hodnoty pro různé veličiny jsou založeny na měřeních v odlišných časech.



Obrázek 4. Náhrada chybějící hodnoty posunutím

Jinou možností je vynechat hodnotu jak u atributu A tak i u atributu B. V tomto případě si okna odpovídají, ale ztratili jsme část naměřených hodnot. V našem případě by to znamenalo, že využijeme pouze kompletní kontroly. Všechny kontroly, při kterých nebyla zjištěna alespoň jedna veličina ze sledovaného souboru, bychom vynechali. Je zřejmé, že tato metoda je využitelná pouze pro problémy, kde chybí jen velmi málo hodnot.

3.2 Náhrada chybějící hodnoty novou hodnotou

Další možností je ponechat značku pro chybějící hodnotu v časové řadě co nejdéle s tím, že v závěru předzpracování dojde k náhradě chybějící hodnoty některým standardním postupem. V datech, kde se v hodnotách měřených atributů projevuje jistá časová „setrvačnost“, je vhodné použít náhradu pomocí průměru hodnot pro sousední časové značky. Toto řešení se sice zpočátku jeví jako příliš složité, umožňuje však na sebe dobře vázat všechny sledované atributy. Situace je znázorněna na obrázku 4.b.

4 Závěr

Úpravami navrženými v odstavci 3.2 se podaří získat data, ve kterých je možné dobře sledovat souvislost mezi vývojovými trendy různých měřených atributů. Těmto úvahám bude věnována jiná samostatná studie, první analýzy jsou obsaženy v [5].

Výše uvedené úvahy a jednoduché postupy mohou vést k přesvědčení, že identifikace anachronických atributů je triviální záležitostí. Je třeba si však uvědomit, že na počátku úlohy dolování dat má hluboké znalosti o úloze pouze expert z oblasti úlohy, který však má jen mlhavou představu o procesu dolování dat (například nezná význam pojmu anachronický atribut). Navíc při předzpracování dat nemusí být jasně definován cíl úlohy a tím ani kritické vazby mezi atributy. Ani v tomto článku pořadí kapitol neodpovídá skutečnému pořadí kroků v čase, zmíněný anachronický atribut byl objeven až ve fázi hodnocení jednoho z prvních vytvořených modelů.

Pro identifikaci anachronických atributů má velký význam porozumění datům podpořené důkladnou analýzou dat. Při práci s časovými řadami, které obsahují různorodý počet časových záznamů se nám osvědčil následující postup:

1. Analýza vztahů mezi cílovou klasifikací a počtem časových záznamů v každém prvku trénovací množiny - tento krok může upozornit na anachronicitu.
2. Volba délky okna pomocí analýzy počtu časových záznamů v jednotlivých případech.
3. Transformace dat pomocí klouzavého okna.

Úprava dat pomocí klouzavého okna bude zakomponována do nástroje pro předzpracování dat SumatraTT [2]. Nástroj nabízí široké možnosti předzpracování dat a je volně dostupný na stránkách [9].

Annotation. The paper is concerned with mining data collected from a number of objects during repeated control measurements all of which are tagged by the corresponding time. No attribute-valued machine learning algorithm can be applied directly on such data provided that the number of controls is not fixed but if it varies. The available data have to be transformed and preprocessed in such a way that uniform type of information is obtained about all the considered objects. This can be achieved e.g., by aggregation. But this process can bring in anachronistic variables, i.e., variables containing information which is not actually available in the data when a prediction is needed. The paper suggests a method how to preprocess considered type of data without falling into the trap of introducing anachronistic attributes. The method is illustrated on a case study based on STULONG data.

Reference

1. Antunes, C. M., Oliveira, A. L.: *Temporal Data Mining: An Overview.* Workshop on Temporal Data Mining, 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001), 2001.
2. Aubrecht, P., Železný, F., Mikšovský, P., Štěpánková, O.: *SumatraTT: Towards a Universal Data Preprocessor.* Proc. of the 16th European Meeting on Cybernetics and Systems Research - Vienna 2002, pp. 818-823, Austrian Society for Cybernetic Study.
3. Baxter, R. A., Williams, G. J., He, H.: *Feature Selection for Temporal Health Records.* Fifth Asia-Pacific Conference on Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, Springer, 2035, pp. 198-209, 2001.
4. Mařík, V., Štěpánková, O., Lažanský, J. a kol.: *Umělá inteligence 4.* Academia, Praha, s. 355-407, 2003.
5. Novakova, L., Klema, J., Jakob, M., Rawles, S., Stepankova, O.: *Trend analysis and risk identification* Discovery Challenge, ECML/PKDD2003, Dubrovnik Croatia 2003, http://euromise.vse.cz/stulong/publikace
6. Pyle, D.: *Data Preparation For Data Mining.* Morgan Kaufmann, California, 1999.
7. Skalská, H.: *Odhady AUC a jejich testy významnosti.* Znalosti 2003, Ostrava, s. 163-170, 2003.
8. Projekt STULONG, WWW page, http://euromise.vse.cz/stulong.
9. SumatraTT, WWW homepage, http://krizik.felk.cvut.cz/Sumatra