Scalers effect on performance of standard Machine learning models

Marek Ciklamini [1]

[1] ***CTU FME, Praha 6, Technická 4, (Marek.Ciklamini@fs.cvut.cz)***

Abstrakt: Modely strojového učení (ML) mohou být často užitečné při zpracování různorodých dat z širokého spektra rozličných odvětví. Díky charakteru empiricky získaných dat jsou ML techniky využívané tam, kde exaktní přístupy matematického modelování nelze jednoduše použít. Tento text navrhuje základní kroky pro zvolení vhodného klasifikátoru určeného k vyhodnocování mnoharozměrného souboru dat.
Klíčová slova: škálování dat, dataset, modely strojového učení, přesnost modelu

Abstract: Machine learning models (ML) might be very useful for postprocessing various of dataset from different scientific fields. The empirical character of mined data is usually obstacle to create exact approach of mathematical model. Text suggests basic steps in order to achive suitable classificator for evaluation of multdimensional dataset.
Keywords: scalers, dataset, machine learning models, model accuracy

1 Úvod

Pro ilustrativní účely a ulehčení popsání navrhovaného doporučeni je vybrán soubor dat (DS) z lékařské oblasti. Heart disease DS [1] (soubor dat zabývající se srdečním onemocněním) je veřejně dostupný a často prezentovaný DS, který obsahuje v redukované podobě 14 atributů a to např. věk pacienta, pohlaví, zbytkový cholesterol, atd. a obsahuje především výstupní informaci o stavu srdečního onemocnění a to pro 303 pacientů.

1.1 Jak lze hledat vhodný ML model?

1. Sestavení korelační matice
2. Definování metod škálování
3. Selekce standardních ML modelů
4. Vyhodnocování základních metrik

1.2 Korelační matice

či taktéž pod pojmem teplotní mapa je znám prvotní indikátor, který ukazuje, zda-li je vůbec smysluplné použití standardních ML technik, a to tehdy pokud jsou přítomny mezi atributy přinejmenším náznaky korelací (hodnoty > abs(0.2)). Nejsou-li přítomné, často selhávají, i jiné metody umělé inteligence, jako je hlubokého učení a jiné.

Obr. 1: Korelační matice DS

1.3 Metody vícerozměrného škálování dat (Scalers)

jsou vybrány pro jejich dostupnost v knihovně Scikit-learn [2], která je dostupná v Python jazyku.

- Standard Scaler: standardizování odstraněním střední hodnoty a škálování k rozptylu jednotlivého atributu
- MinMax Scaler: transformování vektorů pomocí jejich extremních hodnot do daného rozsahu, např. <0,1>.

- MaxAbs Scaler: vektory jsou škálovány k maximu jejich absolutních hodnot taktéž do daného intervalu
- Robust Scaler: použití statistické metody robustní k odlehlým bodům.
- Quantile transform: transformace odhadnutím kumulativní distribuce daného parametru tak, aby měl výsledný vektor požadované rozdělení (Gaussovo, či rovnoměrné).

1.4 Typické klasifikátory

Oblast strojového učení je rok od roku rozsáhlejší a je mimo rozsah článku vyjmenovat a využít všechny dostupné metody. Pro demonstrování přístupu jsou vybrány následující známé metody matematické statistiky [2]:

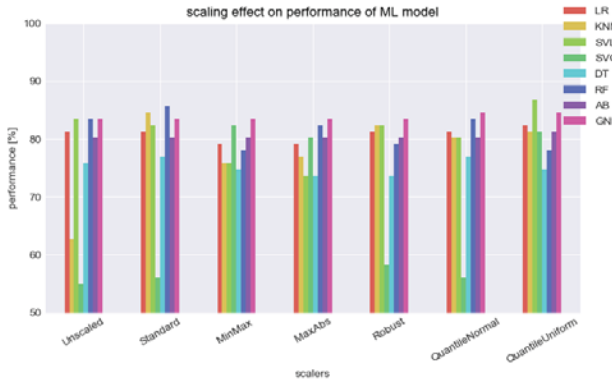
- LR ... Logistická regrese
- KNN ... Nejbližší sousedé
- SVL ... Podpůrné vektory s lineárním jádrem
- SVG ... podpůrné vekt. s radiální bázovou funkcí
- DT ... Rozhodovací strom
- RF ... Náhodný les
- AB ... Adaptive boosting
- GNB ... Najivní Bayessův klasifikátor

2 Vliv škálování na vlastnosti ML modelů

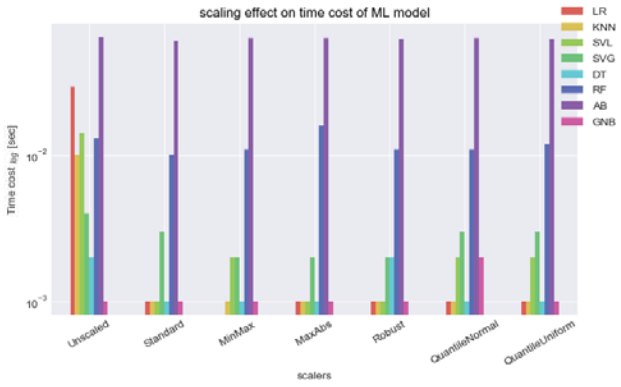
Připravené seznamy metod jsou zpracovány tak, že vznikne škálovaný DS, pro který je vždy vytvořena trénovací/testovací podmnožina v poměru 70/30% a dodatečně je vyžadováno náhodné promíchání vzorků. Jednotlivý typ ML modelu je trénován na škálovaném DS a ten je následně vyhodnocován pro jeho schopnost rozpoznat na testovací podmnožině výstup - pacientův stav.

Přesnost s jakou lze odhadnout v testovací množině výsledek, je nejvyšší pro SVL model a to s 86.8% , tento model je trénován na DS škálovaným Quantilovou transformací, jak lze vidět na Obr. 2

Rychlost trénování ML modelů pro jednotlivé metody škálování je velmi rozdílná tak, jak je patrné na Obr. 3. Hezkým příkladem je model LR trénovaným na původním DS, kdy náročnost je více jak 100 násobná oproti trénovaným modelům pro různě škálovaný DS. Tento aspekt roste na důležitosti s rozměrem DS, kde zde pro demonstrovaný DS je tento aspekt zanedbatelný, avšak pro DS spadající do Big Data (soubor dat mající velikost v GigaByte a více) už je škálování nevyhnutelně zapotřebí.



Obr. 2: Přesnost odhadu



Obr. 3: Rychlost trénování

3 Závěr

Každý model je špatný, avšak některý může být užitečný [3] a pro ML oblast toto tvrzení je opravdu výstižné. V článku byl okrajově demonstrován vliv škálování dat na důležité vlastnosti rozličných ML modelů. Je proto vhodné zdůraznit, že výsledná nejlepší kombinace škálování a modelu lze použít pouze pro daný lékařský DS, kdy na odlišném problému z podobné oblasti (přírodní, lékařské) může být s navrhovaným přístupem dosaženo podobné či stejné kombinace, avšak s nízkou pravděpodobností bude obdobných výsledků dosaženo například pro výrobní data.

Literatura

[1] Matthias Pfisterer Robert Detrano Andras Janosi, William Steinbrunn. Heart disease data set.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] George Box. *Journal of the American Statistical Association*, 1976.

Selected article from
Tento dokument byl publikován ve sborníku

Nové metody a postupy v oblasti přístrojové techniky, automatického řízení a informatiky 2019

New Methods and Practices in the Instrumentation, Automatic Control and Informatics 2019

27. 5. – 29. 5. 2019, Zvíkovské Podhradí

ISBN 978-80-01-06617-1

Web page of the original document:
http://iat.fs.cvut.cz/nmp/2019.pdf

Obsah čísla/individual articles:
http://iat.fs.cvut.cz/nmp/2019/

Ústav přístrojové a řídicí techniky, FS ČVUT v Praze, Technická 4, Praha 6