Stabilne i wydajne metody selekcji cech z wykorzystaniem systemów uczących się

Autoreferat

Miron Bartosz Kursa

13 października 2016

1 Wstęp

W większości zagadnień statystycznego modelowania istnieje konieczność poprzedzenia analizy zmianą surowej formy otrzymanych danych, która jest zazwyczaj podyktowana przez sposób ich pozyskania. Proces ten może przybierać różne formy i służyć różnym celom [4]; w najprostszym przypadku składają się na niego zabiegi techniczne wymagane przez dalej stosowane metody, takie jak zmiany kodowania czy normalizacja.

Zmiana formy danych może również służyć wprowadzeniu wiedzy eksperckiej w celu podkreślenia potencjalnie istotnych aspektów. Można na przykład zastosować przekształcenia uwydatniające interesujące zmienności czy też poszukiwać *a priori* znanych wzorców.

Z drugiej strony, może ona również służyć usuwaniu zbędnej informacji lub szumu, czy też po prostu zmniejszeniu fizycznego rozmiaru danych, tak aby uprościć czy wręcz umożliwić zastosowanie bardziej złożonych obliczeniowo metod. Co naturalne, gdy stężenie istotnej informacji w danych rośnie, badany problem może stać się bardziej oczywisty, co za tym idzie prostszy dla modelowania.

Zmiana reprezentacji może być jednak równie dobrze szkodliwa; niesie za sobą ryzyko wyolbrzymienia wpływu fałszywych wzorców czy czysto losowych zależności ponad te faktycznie występujące, a także usunięcia przydatnej informacji czy wytłumienia efektów ekstremalnych i rzadkich. Co za tym idzie, konieczne jest zapewnienie by nowa forma danych była nie tylko optymalna dla metod stosowanych dalej, ale wciąż reprezentatywna dla badanego zjawiska i zachowująca

możliwie dużą część pierwotnej informacji. Poszukiwania odpowiednich do tego celu metod stanowią wyjście dla niniejszej rozprawy.

2 Selekcja cech

W szczególności, w rozprawie zajmuję się *selekcją cech*, specjalnym przypadkiem zmiany reprezentacji, często jednak wykorzystywanym w praktyce [5]. Dotyczy ona danych dających się przedstawić w formie tabelarycznej jako zbiór *obiektów* opisanych przez wartości pewnego zbioru *atrybutów* (nazywanych też *cechami* czy *zmiennymi*), oraz, w przeciwieństwie do bardziej ogólnych podejść zdolnych do tworzenia nowych cech, jest zdefiniowana jako proces wydzielenia ścisłego podzbioru pierwotnych atrybutów. Główną zaletą takiego ograniczenia jest fakt, że pozwala ono zachować bezpośrednie związki między cechami a konkretnymi fizycznymi aspektami badanego zjawiska, co za tym idzie pozwala zmianie formy danych nadać sens wyjaśniający.

Wyróżnia się dwie fundamentalne klasy metod selekcji cech [13]: poszukujących możliwie małego podzbioru cech, zapewniającego możliwie dobrą dokładność jakiejś metody modelowania (*minimal optimal*) oraz poszukujących podzbioru wszystkich cech, które niosą istotną informację i przez to są potencjalnie użyteczne dla dowolnej metody modelowania (*all relevant*).

W rozprawie argumentuję, że mimo iż metody minimal optimal są na tyle bardziej popularne of all relevant, że są często traktowane na równi z całą klasą metod selekcji cech, to ograniczają wyjaśniającą rolę selekcji oraz wprowadzają istotne ryzyko utraty odporności całej analizy na szum i fałszywe, przypadkowe zależności. Tyczy się to szczególnie problemów $\mathfrak{p} \gg \mathfrak{n}$, to jest takich dla których liczba atrybutów jest istotnie większa od liczby obiektów; takie dane są powszechne chociażby jako wyniki wysokoprzepustowych eksperymentów biologicznych, które są obecnie jednym z najważniejszych narzędzi biologii molekularnej i fundamentem rodzącej się zindywidualizowanej medycyny [3]. Tego typu pomiary pozwalają jednocześnie uchwycić aktywność tysięcy czy nawet milionów agentów opisujących reprezentatywną cześć całego stanu danego układu; z drugiej strony, ze względu na koszty i charakterystykę tego takich badań, odnoszą się do kilkudziesięciu, rzadziej kilkuset obiektów.

Co ważne, wybór analizowanych tak aspektów jest w gruncie rzeczy ślepy, co czyni szczególnie istotną wyjaśniającą rolę selekcji cech, jako że na jej podstawie można wykryć ślady nieznanych wcześniej mechanizmów i przeprowadzić kolejne badania w ich kierunku. Jest jasne, że taki wynik może być nawet istotniejszy niż

pierwotny cel uzyskania dobrego modelu predykcyjnego.

Istnieje także techniczny podział metod selekcji cech, oparty o związek selekcji z modelowaniem [15]: mamy tu *filtry*, algorytmy niezależne od zastosowanej metody modelowania, *metody wbudowane*, integrujące modelowania i selekcję w jeden algorytm, w końcu *metody typu wrapper* które używają modelowania, ale jedynie w roli wyroczni orzekającej o jakości danego wyboru atrybutów.

Zasadniczo, wszystkie filtry powinny być metodami all relevant, gdyż nie są związane z modelowaniem; wiele z nich jest jednak oparte o heurystyki związane z charakterystyką konkretnych metod, najczęściej sprowadzające się do eliminacji redundancji między cechami. Takie podejście może więc wprowadzić problemy charakterystyczne dla selekcji minimal optimal. Co więcej, filtry są zazwyczaj ograniczone do analizy trywialnych lub bardzo prostych zależności między atrybutami i decyzją, co często prowadzi do miernych efektów [14], lub też wymagają istotnych nakładów mocy obliczeniowej na przeprowadzanie wyczerpujących przeglądów bardziej złożonych zależności.

Jak już wspomniałem, metody typu wrapper przeglądają przestrzeń wszystkich możliwych selekcji cech przez analizę wyników modelu dopasowanego do danych obciętych do konkretnych wyborów atrybutów. Metody typu wrapper najczęściej poszukują rozwiązań minimalizujących błąd predykcji, co czyni je typowymi podejściami minimal optimal; kilka algorytmów stosuje jednak bardziej złożone kryteria, w szczególności pozwalające na przeprowadzenie selekcji all relevant. Samo przeszukiwanie może być wyczerpujące, prowadzone przez jakiś algorytm optymalizacji albo też wiedzione dodatkowymi informacjami zwrotnymi z modelu. Niezależnie od tego, zazwyczaj wymagane jest zbudowanie bardzo wielu modeli, a często również zoptymalizowanie ich hiperparametrów przy użyciu złożonych metod takich jak walidacja krzyżowa; co za tym idzie metody typu wrapper są zazwyczaj bardzo wymagające obliczeniowo.

Metody wbudowane są najczęściej istotnie wydajniejsze obliczeniowo niż metody typu wrapper, ale są często metodami klasy minimal optimal, na przykład gdy są oparte o regularyzację, albo zwracają tylko ranking cech, muszą więc być wbudowane w bardziej złożony algorytm aby zwrócić ścisłą selekcję.

3 Rezultaty

W zasadniczej części rozprawy prezentuję heurystyczny wrapper klasy all relevant, metodę Boruta. Jest ona oparta o koncepcję rozszerzania systemu informacyjnego o *cienie*, z definicji nieistotne atrybuty, które są wykorzystywane jako odniesie-

nie dla oceny istotności oryginalnych atrybutów w kontekście pełnej struktury analizowanych danych. Porównanie jest oparte o *ważność atrybutów*, miarę przydatności każdej cechy która jest szacowana przez pewne systemy uczące się w czasie treningu.

W tej roli jest domyślnie wykorzystywana metoda lasu losowego (*Random Fotest*) [2], jako że pozwala na analizę złożonych zależności między cechami, stochastycznie dobiera cechy do modelu w związku z czym dobrze szacuje ważność mniej ważnych zmiennych, w końcu jest mocno odporna na przeuczenie i praktycznie nie wymaga optymalizacji hiperparametrów.

Poza wykorzystaniem cieni, metoda Boruta prowadzi selekcję iteracyjnie; z systemu informacyjnego stopniowo usuwane są cechy uznane za nieistotne, pozwalając wykorzystanemu źródłu ważności dokładniej szacować ważność pozostałych atrybutów, co zwiększa stabilność i dokładność całego algorytmu. Metoda Boruta jest opublikowana w pracy [10].

Następnie odnoszę do dużych wymagań obliczeniowych metody Boruta, uniemożliwiającym jej praktyczne zastosowanie w wielu aplikacjach. W tym kontekście przywołuję metodę paproci losowych, głęboko stochastyczny system uczący się wprowadzony jako wydajna obliczeniowo alternatywa dla metody lasu losowego w wymagających zastosowaniach związanych z analizą obrazów [17, 1], i proponuję jej zmodyfikowaną wersję. Wprowadzona metoda, rFerns, w przeciwieństwie do pierwotnej formy paproci losowych jest klasyfikatorem ogólnego przeznaczenia i jest zdolna do szacowania ważności zmiennych. Metoda rFerns jest opublikowana w pracy [7], wraz z oceną jej jakości, wydajności i użyteczności miary ważności zmiennych, opartą o porównanie z metodą lasu losowego.

W dalszej części rozprawy dokonuję wyczerpującej oceny metody Boruta jak i jej odmiany wykorzystującej rFerns jako źródło ważności na serii problemów klasy $p \gg n$ pochodzenia biologicznego. W szczególności, analizuję tu stabilność selekcji; w tym celu proponuję nową metodę jej oceny opartą o samozgodność wyników szacowaną przy pomocy metody bootstrap. Wyniki tych analiz empirycznie potwierdzają że przyjęcie celu selekcji minimal optimal może z łatwością doprowadzić do znaczącej utraty stabilności selekcji. Co więcej, dowodzą też że popularna w podobnych pracach ocena jakości selekcji przez błąd predykcji modelu wytrenowanego tylko na wyselekcjonowanych atrybutach może prowadzić do fałszywych wniosków. Z drugiej strony uzyskane wyniki potwierdzają słuszność przyjętych heurystyk i efektywność zaproponowanych metod klasy all relevant. Rozważania te zawierają się w pracy [8].

Rozprawę uzupełnia studium stosowalności algorytmu paproci losowych w innym kontekście, wydajnego rozpoznawania instrumentów w nagraniach au-

dio. Zagadnieniem tym zajmowałem się już wcześniej [11, 9, 6]; w trakcie tych badań zostało zaproponowane efektywne podejście wykorzystujące metodę lasu losowego, jednakże także tu poprawa wydajności obliczeniowej byłaby istotna w pewnych zastosowaniach, na przykład implementacji dla systemów wbudowanych czy do masowego przeszukiwania obszernych baz danych. W rozprawie pokazuję że nawet trywialna zamiana metody lasu losowego na metodę paproci losowych zapewnia istotne przyśpieszenie, zarówno treningu jak i klasyfikacji, bez zauważalnej różnicy w jakości predykcji. Co więcej, elastyczność metody rFerns pozwoliła mi przeformułować ją w formie klasyfikatora wieloetykietowego, co pozwoliło uzyskać dalszą redukcję wymagań obliczeniowych. Ten wątek składa się na prace [16] oraz [12].

4 Wnioski

W rozprawie argumentuję, że najczęściej spotykana forma selekcji cech, poszukiwanie zbioru atrybutów optymalizujących jakość predykcji finalnego klasyfikatora, jest mocno podatna na przypadkowe zależności, szczególnie w problemach o dużym wymiarze danych, jak również jest skłonna do usuwania istotnej, choć bardziej subtelnej informacji. Co za tym idzie, może prowadzić do powstania obciążonych modeli i utrudnia wnioskowanie na ich podstawie. Podobnych problemów można jednak uniknąć poprzez zastosowanie selekcji cech klasy all relevant, takich jak zaprezentowana w rozprawie metoda Boruta.

Metoda ta jest jednak stosunkowo wymagająca obliczeniowo, co istotnie ogranicza jej stosowalność, szczególnie dla większych zbiorów danych. Jednym z rozwiązań tego problemu może być zastąpienie wymaganego przez metodę źródła ważności zmiennych na równie dokładne ale znacząco wydajniejsze obliczeniowo niż domyślna metoda lasu losowego. W tym celu proponuję metodę rFerns, uogólnienie metody paproci losowych zdolne do analizy ogólnych zbiorów danych oraz dla którego sformułowałem algorytm oceny ważności atrybutów. Pokazuję, że z użyciem metody rFerns można prowadzić selekcję metodą Boruta na obszernych, realistycznych zbiorach danych w rozsądnym czasie.

Pokazuję też stosowalność wspomnianej wersji metody paproci losowych w zagadnieniu wydajnej analizy danych audio.

Literatura

[1] A. Bosch, A. Zisserman, and X. Munoz. Image Classification Using Random Forests and Ferns. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[2] L. Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

[3] E.R. Dougherty. The fundamental role of pattern recognition for gene-expression/microarray data in bioinformatics. *Pattern Recognition*, 38(12):2226–2228, 2005.

[4] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[5] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, New York, 2013.

[6] E. Kubera, M.B. Kursa, W.R. Rudnicki, R. Rudnicki, and A.A. Wieczorkowska. All that Jazz in the Random Forest. In M. Kryszkiewicz, H. Rybiński, A. Skowron, and Z. Raś, editors, *Foundations of Intelligent Systems*, pages 1–10. Berlin Heidelberg, 2011.

[7] M.B. Kursa. rFerns: An implementation of the random ferns method for general-purpose machine learning. *Journal of Statistical Software*, 61(1):1–13, 2014.

[8] M.B. Kursa. Robustness of random forest-based gene selection methods. *BMC Bioinformatics*, 15(1):1–8, 2014.

[9] M.B. Kursa, E. Kubera, W.R. Rudnicki, and A.A. Wieczorkowska. Random Musical Bands Playing in Random Forests. In M. Szczuka, M. Kryszkiewicz, S. Ramanna, R. Jensen, and Q. Hu, editors, *Rough Sets and Current Trends in Computing*, pages 580–589. Berlin, Heidelberg, 2010.

[10] M.B. Kursa and W.R. Rudnicki. Feature Selection with the Boruta Package. *Journal of Statistical Software*, 36(11), 2010.

[11] M.B. Kursa, W.R. Rudnicki, A.A. Wieczorkowska, and A. Kubik-Komar. Musical Instruments in Random Forest. In J. Rauch, Z.W. Raś, P. Berka, and T. Elomaa,

editors, *Foundations of Intelligent Systems*, volume 5722 of *Lecture Notes in Computer Science*, pages 281–290. Springer, Berlin, Heidelberg, 2009.

[12] M.B. Kursa and A.A. Wieczorkowska. Multi-label ferns for efficient recognition of musical instruments in recordings. In T. Andreasen, H. Christiansen, J.-C. Cubero, and Z. W. Raś, editors, *Foundations of Intelligent Systems: 21st International Symposium, ISMIS 2014, Roskilde, Denmark, June 25-27, 2014. Proceedings*, pages 214–223. Springer International Publishing, Cham, 2014.

[13] R. Nilsson, J.M. Peña, J. Björkegren, and J. Tegnér. Consistent feature selection for pattern recognition in polynomial time. *Journal of Machine Learning Research*, 8:612, 2007.

[14] Y. Saeys, T. Abeel, and Y. Van de Peer. Robust Feature Selection Using Ensemble Feature Selection Techniques. In W. Daelemans, B. Goethals, and K. Morik, editors, *Machine Learning and Knowledge Discovery in Databases*, number 5212 in Lecture Notes in Computer Science, pages 313–325. Springer Berlin Heidelberg, 2008.

[15] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–17, October 2007.

[16] A.A. Wieczorkowska and M.B. Kursa. A Comparison of Random Forests and Ferns on Recognition of Instruments in Jazz Recordings. In L. Chen, A. Felfernig, J. Liu, and Z. Raś, editors, *Foundations of Intelligent Systems*, pages 208–217. Springer, Berlin Heidelberg, 2012.

[17] M. Özuysal, M. Calonder, V. Lepetit, and P. Fua. Fast keypoint recognition using random ferns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):448–461, 2010.