Введение

Целью данной работы является рассмотрение роли доверия в применении систем искусственного интеллекта в повседневной жизни; способах повышения интерпретируемости моделей в общем и возможность использования байесовских сетей для создания более прозрачных моделей машинного обучения, тем самым повышая доверие к ним.

Байесовские сети представляют собой вероятностные модели на основе направленных ациклических графов, где вершины отображают состояние набора переменных, а ребра - переходы системы в другое состояние. Таким образом байесовская сеть позволяет изобразить процесс машинного обучения

Проблемы в развитии доверия к искусственному интеллекту

Доверие лежит в основе принятия любой технологии. Его можно определить так: 1) убеждение в том, что объект создан во благо человечества 2) готовность полагаться на технологию в ситуациях связанных с риском для здоровья 3) совокупностью этих факторов[1]

Одно из основных препятствий в развитии доверия к искусственному интеллекту заложено в том, что процесс работы непрозрачный и случайный, своего рода “черный ящик”[2]. Это можно преодолеть с помощью создания моделей, способных объяснить принятие решений, однако возникает эффект снижения точности их вычислений. Более прозрачные модели вычисляют результат медленнее и менее точно[3].

Объяснение принятых решений лежит в основе доверия пользователя к верной работе системы, особенно в ситуации когда пользователь не обладает экспертными знаниями в области машинного обучения.[4] Однако для тех же неопытных пользователей прозрачная система может показаться очевидной и правой, в связи с чем слепое доверие превысит гарантию работоспособности системы, и даже неверная работа не будет поддаваться критической оценке. Такой феномен описан в опыте с использованием “Копировальной машины”[5]

Способы улучшения интерпретируемости моделей

При обработке данных часто приходится работать не только с большим количеством записей, но и количество атрибутов у каждой записи тоже неуклонно растет. Проблема большой размерности данных в том, что сложно адекватно составить геометрическое представление данных в виде графика или кластеров. Для уменьшения размерности данных рассмотрим два подхода: отбор признаков[6](feature selection) и извлечение признаков[7](feature extraction).

Стоит отметить, что в реальной работе применяются простые алгоритмы. Один из самых распространенных алгоритмов для извлечения признаков это метод главных компонент[8](Primary Component Analysis, PCA). Преимущество данного алгоритма в

том, что с его помощью можно преобразовать исходный набор данных во множество собственных векторов ковариационной матрицы, таким образом выделяются наиболее значимые атрибуты данных, и из-за меньшей размерности их легче проецировать на плоскость.

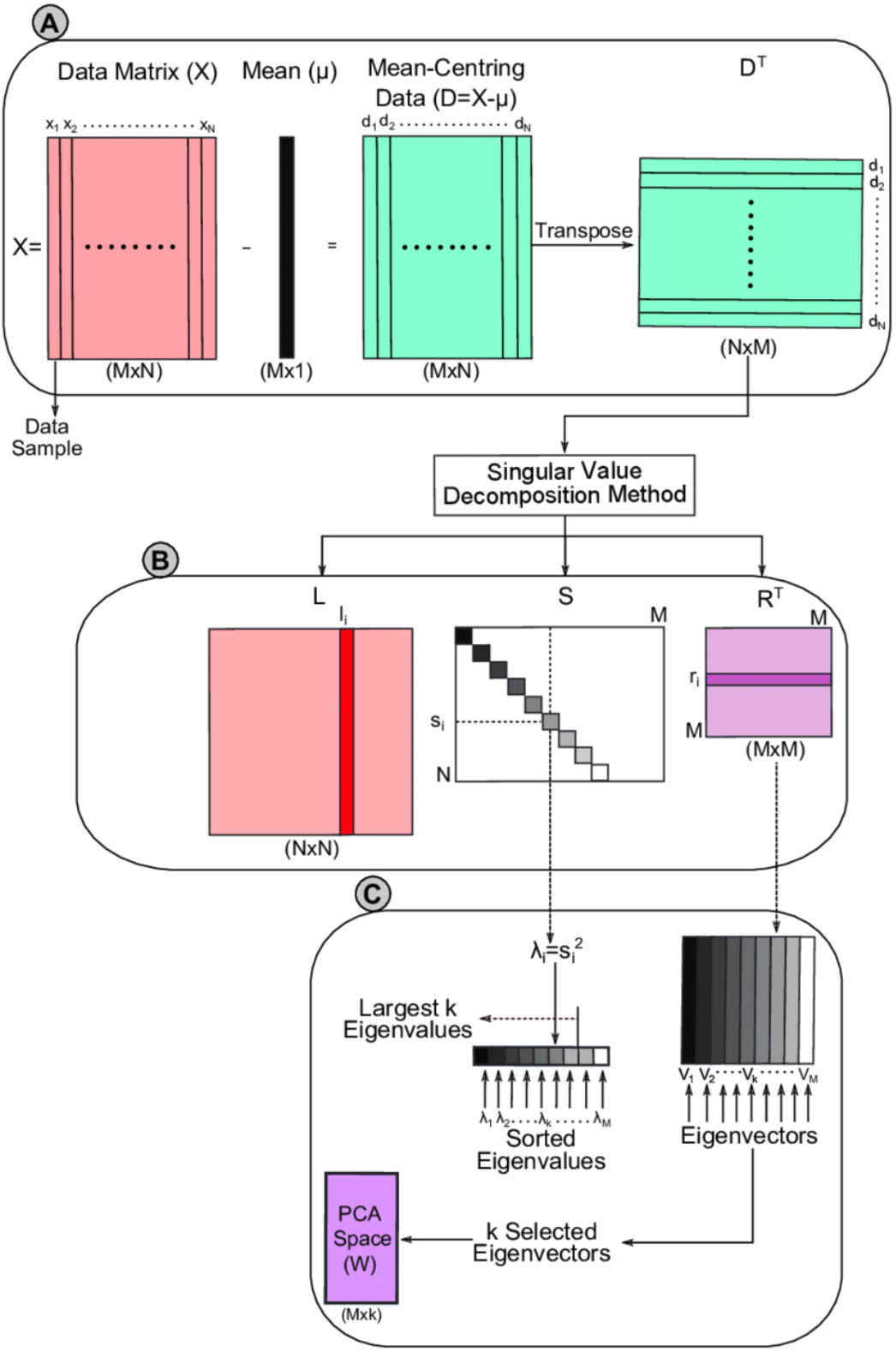

Рисунок 1. Вычисление главных компонент с помощью сингулярного разложения

В опыте применения машинного обучения для классификации МРТ головного мозга[9] рассматривается применение алгоритма отбора признаков(обратный отбор) на однослойном персептроне, с помощью чего удалось достичь приемлемой точности и интерпретируемости модели. В таких случаях использование алгоритмов уменьшения размерности данных практически необходимо, поскольку экспертам нужен не только лишь результат работы, но и качественные, на которых скану был присвоен определенный класс. В данном случае удалось уменьшить размерность данных с 200 параметров до 3.

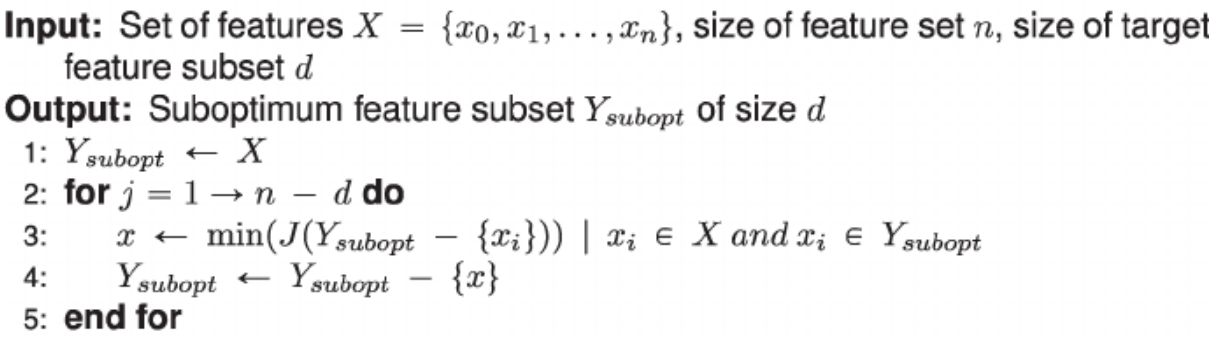

Рисунок 2. Псевдокод алгоритма обратного отбора

Так же существуют алгоритмы уменьшения размерности, результатом работы которых является нелинейная функция. Они слабо распространены из-за сложности проекции данных на плоскость, однако на конференции ESANN регулярно проводятся сессии с докладами по этой теме[10]. Первым был разработан алгоритм Sammon mapping[11], широко применяемый в разведочном анализе данных. Нелинейное разложение применяется в случаях, когда нельзя радикально уменьшить количество параметров данных без потери информации.

Источники

1. Gefen, David. “E-Commerce: The Role of Familiarity and Trust.” Omega, Vol. 28, No. 6, 2000 (https://www.sciencedirect.com/science/article/pii/S0305048300000219)
2. A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on Explainable Artificial Intelligence(XAI)”, IEEE Access, 6, 2018, pp. 52138-52160.
3. A. Holzinger, et al., “What do we need to build explainable AI systems for the medical domain?”, arXiv preprint arXiv1712.09923, 2017.
4. Andras et al., (2018) “Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems”. IEEE Technology and Society Magazine, 37(4), 2018, pp. 6-83.
5. E.J. Langer et al., “The mindlessness of ostensibly thoughtful action: The role of" placebic" information in interpersonal interaction”, Journal of Personality and Social Psychology, 36(6), 1978, pp. 635-642.
6. G. Dy, C.E. Brodley, Feature subset selection and order identification for unsupervised learning. In proceedings of the 17th International Conference on Machine Learning (ICML 2000), Morgan Kaufmann Publishers Inc., pages 247-254, Standford, CA (USA), 2000.
7. I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh (Eds.), Feature Extraction: Foundations and Applications. Studies in Fuzziness and Soft Computing, Springer, 2006
8. I.T. Jolliffe, Principal Component Analysis. Springer; 2nd edition, 2002.
9. A. Vellido, E. Romero, M. Juli`a-Sap´e, C. Maj´os, `A Moreno-Torres, C. Ar´us, Robust discrimination of glioblastomas from metastatic brain tumors on the basis of single-voxel proton MRS. NMR in Biomedicine. Accepted for publication. doi: 10.1002/nbm.1797.
10. A. Wismueller, M. Verleysen, M. Aupetit, J.A. Lee, Recent advances in nonlinear dimensionality reduction, manifold and topological learning. In M. Verleysen, editor, proceedings of the 18th European Symposium on Artificial Neural Networks (ESANN 2010), d-side pub., pages 71-80, Bruges (Belgium), 2010
11. Henderson, Paul. "Sammon mapping." *Pattern Recognit. Lett* 18.11-13 (1997): 1307-1316.