Исследование рекомендательных систем алгоритмов выбора подмножества признаков основанных на мета-обучении

Танфильев И. Д., магистрант кафедры компьютерных технологий Университета ИТМО, tanfilyev@gmal.com

Фильченков А. А., к.ф.-м.н., доцент кафедры компьютерных технологий Университета ИТМО, aaafil@mail.ru

Аннотация

Система мета-обучения рекомендует алгоритм выбора признаков из заданного множества доступных алгоритмов, для данного набора данных. Рекомендованный алгоритм считается оптимальным для данной задачи. Это достижимо с помощью определения наборов данных близких к данному из множества известных системе, используя мета-информацию о таких наборах данных. Задачей данного исследования является исследования различных способов построения рекомендательных систем для достижения наилучшего результата рекомендации.

Введение

Анализ больших объемов данных становится все востребованее, и ученые, как правило, очень ограничены в вычислительных ресурсах. Это ставит вопрос о необходимости применении оптимизаций при работе с данными. Одним из способов выполнить такую оптимизацию в рамках задачи классификации могут служет служить применение алгоритмов выбора подмножества признаков [1, 2]. Правильно подобранный алгоритм может существенно ускорить процесс обработки данных, незначительно сократив точность или даже повысив качество работы классификатора. С другой стороны, ошибка в выборе алгоритма может негативно повлиять на производительность или привести к потере данных.

Стоит отметить, что не существует алгоритма, одинаково хорошо работающего на всех типах данных [3]. Таким образом, возникает задача поиска алгоритма выбора признаков, подходящего для конкретной задачи. Данная проблема не имеет простого решения в силу большого

разнообразия различных алгоритмов выбора признаков, и затрудненности экспертной оценки качества работы этих алгоритмов во многих случаях реального применения.

Целью данной работы является повышение эффективности синтезированного алгоритма выбора признаков, основанного на ранжировании результатов работы рекомендательной системы алгоритмов выбора признаков. Для достижения цели данной работы предлагается использовать различные методы построения рекомендательных систем и исследовать их на различных наборах данных.

Существующие решения в данной области

Подход, описанный в работе [4], позволяет создать рекомендательную систему, основанную на алгоритма классификации kNN. Работа этой рекомендательной системы достаточно хорошо изучена, например, в работе [5], где автор исследовал мета-признаки необходимые для установления схожести между наборами данных. Также эта система исследовалась в работе [6], которая состояла в применнеии алгоритмов ранжирования [7, 8] к результатам работы рекомендательной системы, описанной выше. Также в работе [6] была предложена метрика AEARR, предназначенная для сравнения качества выбранных признаков.

$$\mathrm{AEARR}(S_i) = \frac{1}{M-1} \sum_{j,j \neq i} \frac{F_1(i) + F_1(j)}{1 + \beta \log(|S_i|/|S_J|)}, \tag{1}$$

где S_i — iй множество признаков, M — число множеств участвующих в сравнении, $F_1(i)$ — F_1-мера [9] подсчитанная используя список признаков S_i.

В работе [11] предложены альтернативные способы построения рекомендательных систем, а именно подход бинарной попарной классификации и метод ранжирования меток. Каждый из них достаточно прост и использует классификатор при помощи которого выбираются алгоритмы, подходящие для данной задачи. Такие наборы данных будут содержать мета-признаки уже известных наборов данных, сохраненных в базе данных. В качестве нового объекта будут рассмотрены мета-признаки от исследуемого набора данных.

Подход ранжирования меток заключается в том, что если в системе зарегистрированно N алгоритмов выбора подмножества признаков в набор данных будет добавленно N записей от каждого из наборов данных. Каждая такая запись будет содержать мета-признаки набора

даных, вес данного объекта, вычесленный с помощью AEARR, а также метку в виде названия алгоритма выбора подмножества признаков. Для получения ранжирования в данном методе обрабатывается вероятностное распределение по меткам.

Подход попарной бинарной классификации заключается в том, что для всех возможноых пар алгоритмов выбора подмножества признаков строится наборы данных с бинарной классификацией. Это нужно для того чтобы понять какой из двух алгоритмов выбора признаков окажется лучнше на новом наборе данных. Зная такую информацию обовсех парах можно составить ранжирование алгоритмов, которое является искомым для данного набора данных.

Исследование

Каждая из рекомендательных систем была запущенна на 125 хорошо изученных наборов данных, полученных из различных источников. Для кажного нового набора данных рекомендательная система рекомендовала 5 алгоритмов выбора подмножества признаков. Эти алгоритмы находили наборы признаков для исследуемого набора данных, а затем объединялись при помоши алгоритмов аггрегации ранжирований [7], используя подход описанный в работе [6]. Используя каласификатор на агрегированом множестве признаков строилась оценка AEARR, которая позмоляет оценить и сравнить качество работы рекомендательной системы для конкретного набора данных.

Как мы можем наблюдать из графиков на Рис. 1, классификатор LR-SVN достаточно эффективен.

Также стоит отметить, что все значения достаточно далеки от random. Это означает, что их применение достаточно эффективно в для данной задаче.

При анализе данных, использовался наивный байесовский классификатор [10], который был применен к выбранным признакам для оценки результатов путем построения метрики AEARR. Таеже в при синтезе алгоритмов использовалисть 5 алгоритмов рекомендованных системами мета-обучения. Именно этим объесняется то, что результаты работы алгоритма oracle, иногда, имеют оценку ниже 1.0.

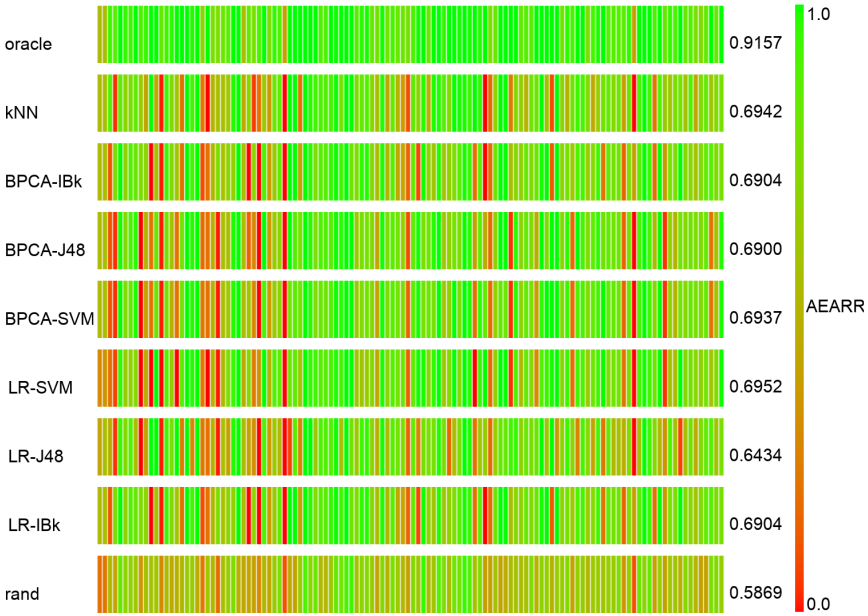

Рис. 1: Сравнение работы классификаторов в зависимости от их типов и заданного алгоритма классификации. На графике oracle - то к чему дожна стремиться система, kNN - сущестующий, подход, BPCA - подход попарного сравнения, LR - подход ранжирования меток, rand - алгоритм который рекомендует случайные алгоритмы.

Заключение

Различные рекомендательные системы хорошо подходят для задачи выбора алгоритмов выбора подмножества признаков и могут успешно применяться к различным наборам данных. Большинство алгоритмов них показывает результаты достаточно близкие к хорошо изученному подходу kNN. Особенно хорошо себя зарекомендовал подход ранжирование меток использующий клссефикатор SVM который превзошел подход kNN и не требует дополнительных затрат времени и ресурсов для обработки.

Также стоит отметить, что большинство рекомендательных систем ошибаются на одних и тех-же данных. Это связанно с тем что на этих данных модель, в том числе и используемые мета-признаки наборов данных, не работает должным образом.

В дальнейшем планируется обратить внимание на то, что многие алгоритмы выбора подмножества признаков выбирают очень схожие наборы признаков. То есть после агригации таких наборов ничего не измениться и качество работы алгоритма не улучшется. Таким образом в дальнейшем стоит обратить внимание на способ выбора алгоритмов выбора подмножества признаков работающий не похожим образом.

Литература

[1] Cateni S., Vannuci M., Vannocci M., Colla V. Variable Selection and Feature Extraction Through Artificial Intelligence Techniques // Multivariate Analysis in Management, Engineering and the Sciences. 2012. Pp. 103–118.

[2] Guyon I., Elisseeff A. An introduction to variable and feature selection // The Journal of Machine Learning Research. 2013. No. 3. Pp. 1157–1182.

[3] Wolpert D.H. Macready, W.G. No Free Lunch Theorems for Optimization // IEEE Transactions on Evolutionary Computation. 1997. Vol. 1, no. 1. Pp. 67–82.

[4] Guangtao W., Qinbao S., Heli S., Xueying Z., Baowen X., Yuming Z. A Feature Subset Selection Algorithm Automatic Recommendation Method // Journal of Artifficial Intelligence Research, 2013. No. 47. Pp. 1–34.

[5] Filchenkov A., Pendryak A. Datasets Meta-Feature Description for Recommending Feature Selection Algorithm // Proceedings of the AINL-ISMW FRUCT - 2015, pp. 11–18

[6] Танфильев И.Д., Сметаников И.Б. Агрегирование ранжирований результатов в задаче выбора подмножества атрибутов на основе мета-обучения // XVIII Международная конференция по мягким вычислениям и измерениям SCM'15 (19-21 мая 2015 г., Санкт-Петербург.). Сборник докладов. Т. 1. СПб: ЛЭТИ. 2015. С. 91–94.

[7] Shili L. Rank aggregation methods // Wiley Interdisciplinary Reviews: Computational Statistics. 2010. Vol. 2, no. 5. Pp. 555–570.

[8] Zabashta A., Smetannikov I., Filchenkov A. Study on Meta-Learning Approach Application in Rank Aggregation Algorithm Selection // MetaSel@PKDD/ECML, 2015. Pp. 115-116.

[9] Van Rijsbergen C. J. Information Retrieval // Butterworth 2nd ed., 1979. P. 147.

[10] George H, Langley P. Estimating Continuous Distributions in Bayesian Classifiers // Eleventh Conference on Uncertainty in Artificial Intelligence. San Mateo. 1995. Pp. 338-345.

[11] Sun, Q. (2014). Meta-Learning and the Full Model Selection Problem. Unpublished PhD thesis // University of Waikato