Методы оптимизации работы алгоритма выбора подмножества признаков на основе рекомендательной системы алгоритмов выбора подмножества признаков и агрегации ранжирований

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

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 улучшение работы рекомендованного алгоритма выбора признаков, путем дополнительной обработки результатов работы алгоритма преобразованием множества признаков в ранжирования.

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 ранжировать признаки, выбранные фильтрующими алгоритмами выбора признаков.

Существующие решения в данной области

Подход, описанный в работе [4], позволяет создать рекомендательную систему, основанную на предположении, что алгоритмы выбора атрибутов, эффективно снижающие размерность данных на одном наборе данных, будет хорошо работать на близких к данному набору данных.

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.

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

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.

Результаты работы системы показали значительный прирост эффективности выбранного списка признаков, по сравнению не агрегированными результатами работы рекомендательной системы.

Данная система содержит один недостаток: большинство использованных в предыдущих работах [4, 6] алгоритмов выбора признаков является фильтрующими, а именно таких алгоритмов 18, тогда как всего в работе рассматривалось 22 алгоритма. То есть результатом работы таких алгоритмов являются не ранжированные списки, а множества признаков. Это означает, что применять алгоритмы агрегации ранжирований некорректно. Для того, чтобы решить эту проблему можно для каждого из множества признаков построить эквивалентное ему ранжирование.

Оптимизации

Для повышения эффективности системы, описанной выше, в данной работе предлагается оценить признаки в каждом из множеств и построить соответствующие им ранжирования. А именно для достижения этой цели рассмотрим следующий алгоритм:

1. Для каждого атрибута из набора данных построить набор без этого атрибута.
2. Оценить набор признаков, построенный на предыдущем шаге, запустив классификатор, используя метрику AEARR, тем самым сопоставив каждому атрибуту его вклад.
3. Для построения итогового ранжирования нужно отсортировать признаки по возрастанию значения метрики AEARR, полученной на предыдущем шаге.

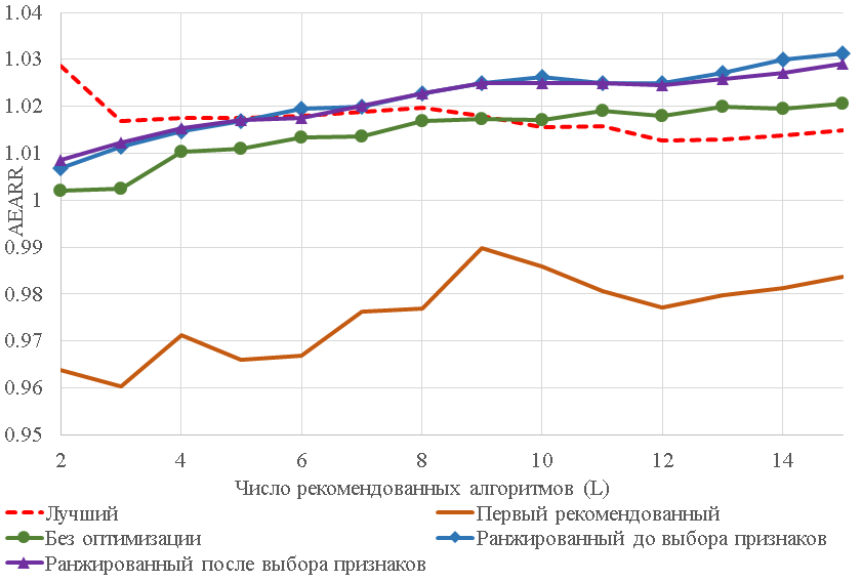

Рис. 1: Эффективность работы системы используя наивный байесовский классификатор [10] при различном заначении параметра L при фиксированном $\beta = 0.1$

Заметим, что такой алгоритм можно применить как до запуска алгоритмов выбора признаков, так и после.

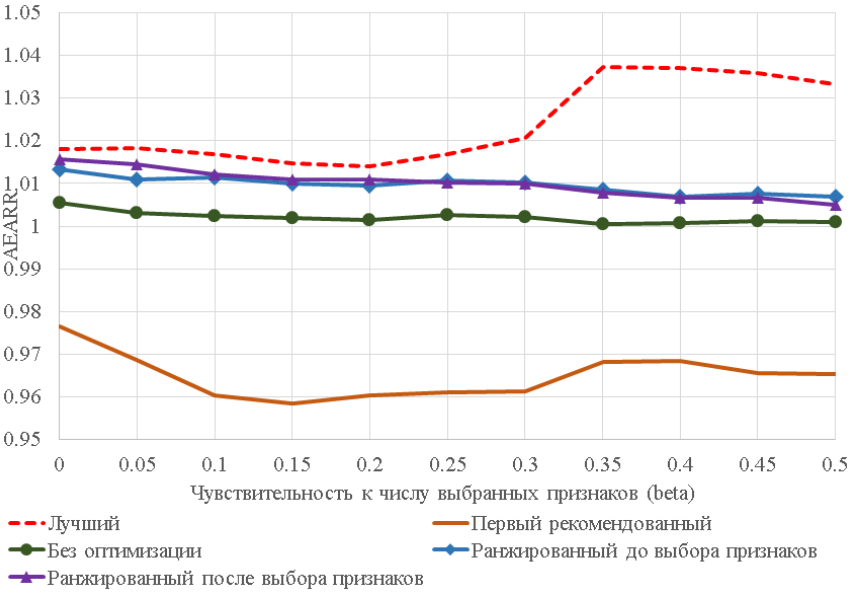

Рис. 2: Эффективность работы системы используя наивный байесовский классификатор [10] при различном заначении параметра β при фиксированном $L = 3$

Как мы можем наблюдать из графиков на Рис. 1 и Рис. 2, наблюдается заметный прирост эффективности синтезированного алгоритма признаков.

При анализе данных, полученных при использовании других классификаторов, таких как J48 [11], PART [12], Байесовская сеть [13] и IBk [14], выяснилось, что ранжирование уже выбранных признаков, в среднем, является более эффективным.

Заключение

Использование дополнительного шага: ранжирования множества выбранных признаков перед агрегацией, показало свою эффективность. Построенная система может успешно применяться к различным наборам данных.

Применение ранжирования признаков, выбранных алгоритмами выбора признаков, оказалось наиболее эффективным в сравнении с предобработанными признаками.

В дальнейшем планируется обратить внимание на то, что классификаторы, обученные на различных множествах признаков, выбранных разными алгоритмами выбора признаков, будут работать с разной эффективностью на различных данных в пределах одного набора данных, что приводит к мысли, что объединение классификаторов, одинаково работающих на одних и тех же данных, возможно, не является эффективным подходом.

Решением описанной проблемы может служить подход, заключающийся в выборе классификаторов, которые работают непохожим образом, и агрегировать соответствующие им списки признаков для получения наиболее эффективного множества признаков. Есть несколько подходов для определения схожести классификаторов. Одним их них является построение кластеризации. Кроме того, возможно построить метрическое пространство похожести классификаторов. Следующим шагом предлагается выбирать случайным образом непохожие классификаторы, для их последующего объединения. При этом появляется возможность использовать только эффективные классификаторы.

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] Quinlan R. C4.5: Programs for Machine Learning // Morgan Kaufmann Publishers, San Mateo, CA. 1993.

[12] Frank E., Ian H. Generating Accurate Rule Sets Without Global Optimization. In: Fifteenth International Conference on Machine Learning. 1998. Pp. 144-151.

[13] Heckerman D. A Tutorial on Learning With Bayesian Networks. 1995.

[14] Aha D., Kibler D. Instance-based learning algorithms // Machine Learning. 1991. No 6. Pp. 37-66.