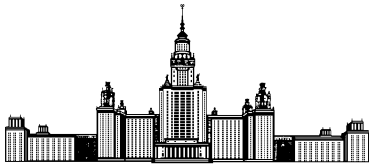
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТКИ 317 ГРУППЫ

«Методы отбора признаков»

Выполнила:

студентка 3 курса 317 группы

Рысьмятова Анастасия Александровна

Научный руководитель:

д.ф-м.н., профессор

Дьяконов Алксандр Геннадьевич

Москва, 2015

Содержание

1 Введение

На этапах постановки задачи машинного обучения и формирования данных не всегда понятно, какие признаки важны для построения оптимального алгоритма, поэтому часто в данных встречается много избыточной (шумовой) информации. Появление шумовых признаков ухудшает качество работы алгоритма и замедляет его работу. Поэтому в большинстве случаев перед решением задачи классификации, регресии или прогнозирования необходимо выбрать те признаки, которые наиболее информативны. Правильный выбор признаков может быть более значимой задачей, чем уменьшение времени обработки данных, или улучшения точности классификации. К примеру, в медицине[1], нахождение минимального набора признаков, который является оптимальным для задачи классификации, может быть полезным для разработки диагностического теста.

Отбор важных признаков (таких как гены, соответствующие для определенного типа рака) может помочь расшифровать механизмы, лежащие в основе проблемы, представляющей интерес для исследования.

В данной работе представлены основные методы отбора признаков, а также приведен пример задачи, при решении которой один из методов показал хороший результат, рассмотрен пример работы некоторых из алгоритмов отбора признаков для задачи определения факторов влияющих на стоимость арендной платы нежилых помещений в США.

2 Методы отбора признаков

Метод отбора может быть реализован с помощью полного перебора признаков. Такой метод прост в реализации, но он совершенно неэффективен на больших данных, поэтому в этом случае чаще всего используются другие алгоритмы.

Существуют три основных класса алгоритмов выбора компонентов - *фильтры, обертки* и *встроенные алгоритмы* [4].

2.1 Фильтры

Фильтры основаны на некоторых показателях, которые не зависит от метода классификации. Например такие как, корелляция между признаками. Они применяются до классификации.

Одним из его приемуществ является то, что фильтрация может быть использована в качестве предварительной обработки для уменьшения размерности пространства и преодоления переобучения. К сожалению такие методы не предназначены для обнаружения сложных связей между признаками, и, как правило, не являются достаточно чувствительными для выявления всех зависимостей в данных.

2.2 Встроенные алгоритмы

Встроенные алгоритмы выполняют отбор компонент во время процедуры обучения классификатора, и именно они явно оптимизируют набор используемых признаков для достижения лучшей точности [10]. Приемущества встроенных алгоритмов в том, что как правило они находят решения быстрее, избегая переподготовки данных с нуля, при этом пропадает необходимость разделять данные на обучающую и тестовую подвыборку. Вместе с тем науке не известны какие-либо встроенные методы, позволяющие решить все существующие задачи.

2.3 Методы обертки

Методы обертки опираются на информацию о важности признаков полученную от некоторых методов классификации или регрессии, и поэтому могут находить более глубокие закономерности в данных, чем фильтры. Обертки могут использовать любой классификатор, который определяет степень важности признаков. Подробнее несколько алгоритмов обертки будут рассмотренны в этой работе далее.

3 Алгоритмы отбора признаков с помощью random forest

Random forest — алгоритм машинного обучения, представляет собой ансамбль многочисленных несмещенных, но чувствительныех к обучающеий выборке алгоритмов (деревьев решений). Каждый из этих классификаторов строится на случайном подмножестве объектов и случайном подмножетсве признаков.

Алгоритм RandomForest может быть использован в задаче оценки важности признаков. Кроме того, случайный лес имеет еще некотороые приемущества для использования его в качестве алгоритма отбора признаков: он имеет очень мало настраиваемых параметров, относительно быстро и эффективно работает, что позволяет находить информативность признаков без значительных вычислительных затрат.

3.1 Boruta

Эвристический алгоритм отбора значимых признаков, основанный на использовании Random Forest [5]. Суть алгоритма заключается в том, что копируются признаки, а затем каждый новый признак заполняется случайным образом, путем перетасовки его значений. На полученной выборке запускается Random Forest.

В целях получения статистически значимых результатов эта процедура повторяется несколько раз, переменные генерируются независимо на каждой итерации.

Запишем пошагово алгоритм Boruta :

1. Добавить в данные копии всех признаков. В дальнейшем копии будем называть скрытыми признаками.

2. Случайным образом перемешать скрытые признаки.

3. Запустить random forest и получить Z-меру всех признаков. Z - мера это такая мера, которая считается как средняя потеря деленая на стандартное отклонение

4. Найти максимальную Z-меру из всех Z-мер для скрытых признаков.

5. Удалить признаки, у которых Z-мера меньше чем найденная на предыдущем шаге.

6. Удалить все скрытые признаки.

7. Повторять все шаги до тех пор пока Z- мера всех признака не станет больше чем максимальная Z-мера скрытых признаков

3.2 ACE

ACE (Artificial Contrasts with Ensembles)[6] - еще один алгоритм, который может быть использован для отбора признаков. Главная идея алгоритма ACE схожа с идеей алгоритма Boruta - каждый признак заполняется случайным образом, путем перетасовки его значений. На полученной выборке запускается Random Forest. Однако, в нем, в отличие от Boruta, не удаляются найденные признаки с наименьшей важностью, которые позволяют повысить качество измерений важных признаков. Наиболее важные признакие найденные алгоритмом ACE, наоборот удаляют, что позволяет алгоритму находить более тонкие связи в алгоритмах.

4 Пример использования алгоритмов отбора признаков

4.1 Gene expression data

В статье [1] был рассмотрен пример работы алгоритма Boruta для задачи выявления различия между двумя подтипами лейкоза. Эта задача решалась до этого

другими методами [7], что позволо сравнить результаты после использования Boruta. Данные имеют информацию о 38 больных. По каждому из которых описано 3051 генов. Для оценки качества алгоритма было добавлено еще 1000 полу-синтетических признаков, которые были сгенерированны путем перетасовки 1000 выбранных случайно имеющихся генов. Сгенерированные признаки хороший алгоритм не будет выбирать, как важные.

Значимость гена оценивалась с помощью алгоритма Boruta на основе Random Forest, состоящего из растущего числа деревьев. Число деревьев в Random Forest варьировалось от 500 до 100 000. Каждый запуск повторили 15 раз.

Введем следующие обозначения:

$Dud2002$ - выделенные гены в 2002 году , решение описано в статье [9]

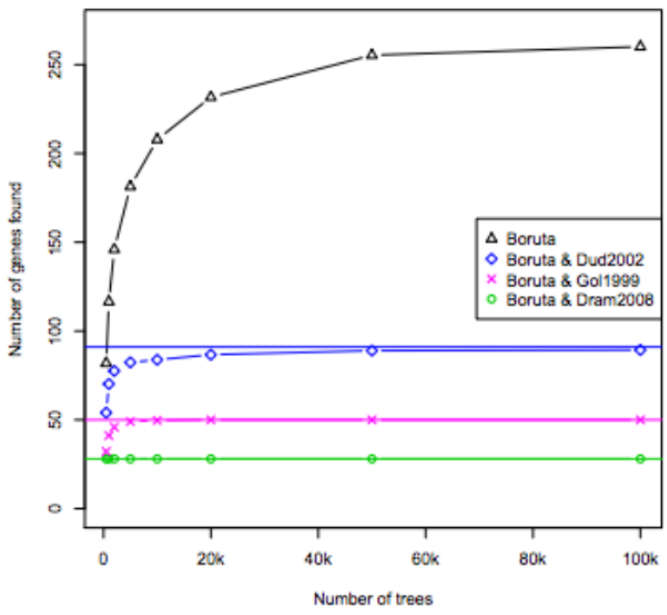
$Gol1999$ - выделенные гены в 1999 году , решение описано в статье [7]

$Dram2008$ - выделенные гены в 2008 году , решение описано в статье [8]

$Bor500$ - выделенные гены с помощью Boruta на основе random forest с 500 деревьями

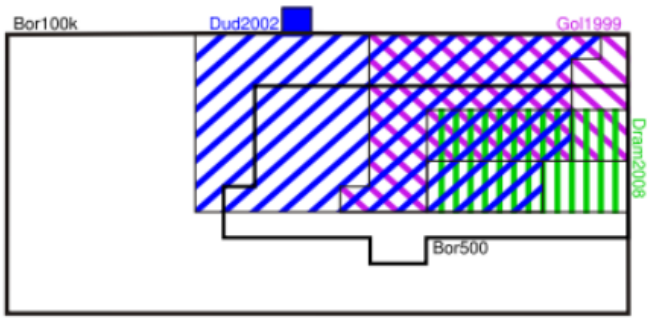
$Bor100k$ - выделенные гены с помощью Boruta на основе random forest с 100000 деревьями

Ниже приведен график зависимости количества выбранных генов, от числа деревьев. Черными точками выделен алгоритм Boruta, а остальными цветами показан результат работы алгоритмов $Dud2002$, $Gol1999$, $Dram2008$, которым на вход подавались признаки выделенные с помощью Boruta.



Видно, что с увеличением числа деревьев увеличивается количество выбранных генов. Сплошные горизонтальные прямые обозначают общее количество признаков выделенных данными алгоритмами.

В результате получилось, что алгоритм ни разу не выбрал в качестве важных генов полу-синтетические данные, добавленные для проверки. На основе этого, авторы данного эксперимента сделали вывод, что алгорим Boruta эффективно справляется с задачей отбора признаков.
Ниже приведено изображение, на котором показано, как пересекаются выбранные множества важных генов у различных алгоритмов используемых разными люди в разное время.



Видно, что множество геннов выделенное с помощью Boruta с 500 деревьюми пересекается со всеми выделенными генами с помощью результатов полученных ранее, а полученные ранее результаты слабо кореллировали между собой. Гены выделенные с помощью Boruta с 100000 деревьями включают в себя все гены выделенные остальными алгоритмами. Данный эксперемент подтверждает, что все выделенные ранее признаки, действительно имеют значение. Более того, алгоритм выделил 150 новых генов.

Результат данного эксперимента показывает, что чувствительность алгоритма Boruta зависит от количества деревьев используемых в random forest. Это происходит благодаря свойству меры важности. Она оценивается как среднее снижение точности деревьев, которые используют данный признак. После этот признак удаляется, поэтому актуальность гена можно узнать только при достаточно большом количестве деревьев.

5 Эксперименты с данными

5.1 Данные

Для изучения работы алгоритмов отбора признаков будут использованы данные о 2227 объектах недвижимости различного типа.

На основе предоставленных данных нужно предсказать стоимость арендной платы и определить признаки от которых наиболее сильно зависит арендная плата. Данные имеют 19 признаков, из них:

2 категориальных:

SpaceType - Тип здания в котором расположен объект, имеет 31 значение

LeaseType - Несёт информацию о том какие расходы включены в арендную плату

9 целочисленных :

SpaceSize - Размер помещения

Number of transport spots - Количесво мест для стоянки транспорта

Population - Количество населения в данном регионе

Landarea - Площадь региона

Social chat score - Престижность (социальный балл)

Average HH income 2013 - Средний доход населения за 2013 год

Average salary of employees($000s) - Средний даход рабочего

Average salary of employees in new businesses - Средний даход рабочего в новом бизнесе

Number of new retail places $2013 - 2010$ - Число новых мест аренды за 2010-2013 год

8 вещественных:

Population change $2013 - 2010$ - Изменение числа населения в процентах

Density of people living in area - Плотность людей проживающих в регионе

Density of people working in area (*based on lat/lon*) - Плотность работающих людей

Total density ($living + working$) - Сумма предыдущих двух признаков

Household size - Размер складского помещения

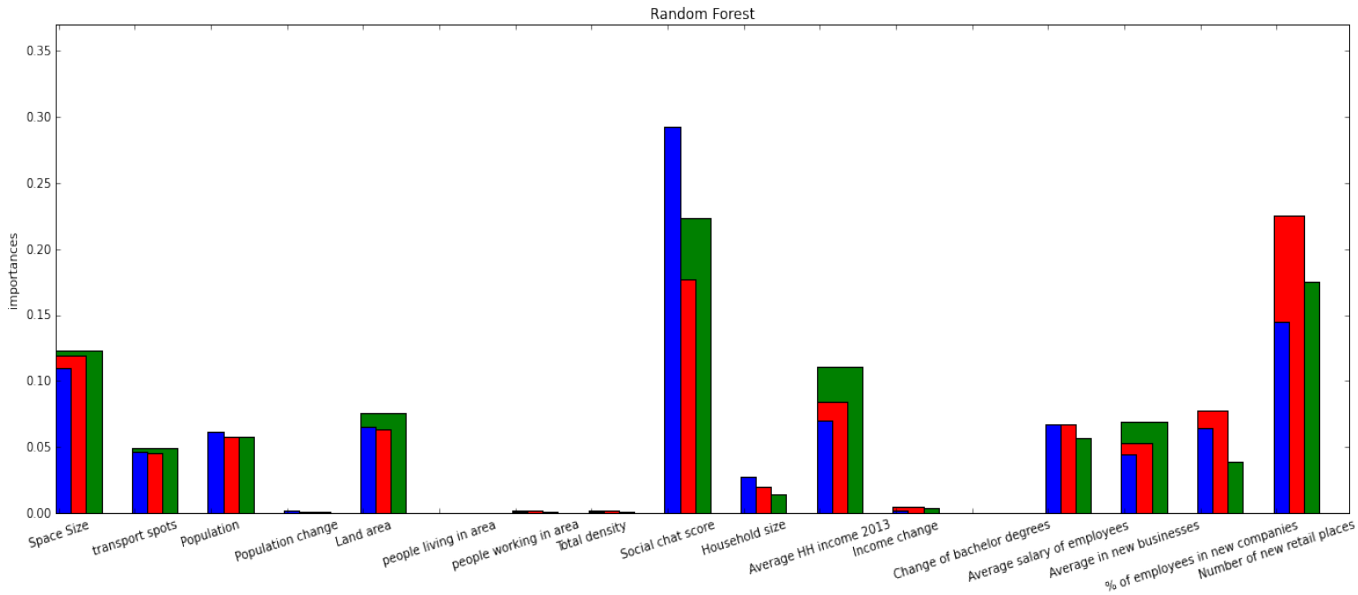
Income change 2013 – 2010 - Изменение дохода за 2010 - 2013 год

Change in % of bachelor degrees - Изменение в процентах количества людей со степенью бакалавра

% of employees in new companies vs all - Процент содрудников в новых компаниях

5.2 Оценка важности признаков

Изначальная выборка была разбита на три части, для использования кросс-валидации по трем фолдам. Ниже приведен график в котором для каждого из 17 некатегориальных признаков показана их важность для трех частей выборки с помощью встроенного алгоритма в random forest



На графике разными цветами указана важность признаков полученная по трем частям выборки.

Максимальную важность в результате имеют следующии признаки:

Social chat score - Престижность (социальный балл)

Number of new retail places 2013 – 2010 - Число новых мест аренды за 2010-2013 год

SpaceSize - Размер помещения

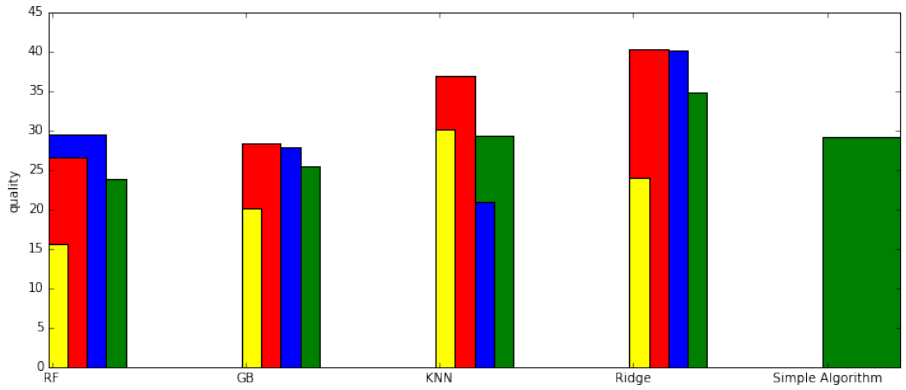
Average HH income 2013 - Средний доход населения за 2013 год

Landarea - Площадь региона

Остальные признаки имели важность в среднем меньшую 0.05. Используя лишь те признаки, которые имели максимальную важность был построен простой алгоритм, средняя ошибка которого по метрике MSE которого мало отличалась от random forest.

После многочисленных экспериментов получилось, что random forest это лучший алгоритм машинного обучения для решения задачи предсказания стоимости арендной платы.

Для задачи использовались и другие алгоритмы машинного обучения. Ниже, на графике приведены результаты тех алгоритмов, которые показали наиболее лучший результат.



На графике указана ошибка алгоритмов по метрике MSE для каждого из трех фолдов и средняя ошибка по трем фолдам.

Так как простой алгоритм, основанный лишь на наиболее важных признаках, дал сравнительно неплохой результат, то можно сделать вывод, о том что встроенный в random forest алгоритм действительно хорошо справляется со своей задачей.

6 Заключение

Отбор признаков является важным этапом построения алгоритмов машинного обучения. Данный этап небходим, чтобы избавиться от шумовых признаков и благодаря этому улучшить качество и ускорить работу алгоритмов. Проведенные экспе-

рименты подтверждают, что алгоритмы отбора признаков с помощью random forest эффективно справляется со своей задачей

Список литературы

[1] The all relevant feature selection using random forest MB Kursa, WR Rudnicki arXiv preprint arXiv:1106.5112 (2011)

[2] Воронцов К. В.: Лекции по методам оценивания и выбора моделей (2007)

[3] Nilsson, R., Pena, J.M., Bjorkegren, J., Tegner, J.: Consistent feature selection for pattern recognition in polynomial time. The Journal of Machine Learning Research 8, 612 (2007)

[4] Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3, 1157–1182 (2003)

[5] Kursa, M.B., Rudnicki, W.R.: Feature Selection with the Boruta Package. Journal Of Statistical Software 36(11) (2010)

[6] Tuv, E., Borisov, A., Runger, G., Torkkola, K.: Feature selection with ensembles, artificial variables, and redundancy elimination. The Journal of Machine Learning Research 10, 1341–1366 (2009)

[7] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (New York, N.Y.) 286(5439), 531–7 (Oct 1999)

[8] Draminski, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., Komorowski, J.: Monte Carlo feature selection for supervised classification. Bioinformatics 24(1), 110–117 (Nov 2008)

[9] Dudoit, S., Popper-Shaffer, J., Boldrick, J.C.: Multiple Hypothesis Testing in Microarray Experiments (2002)

[10] Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3, 1157–1182 (2003)