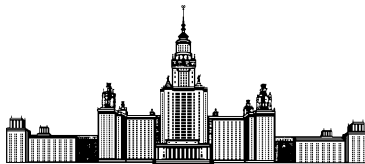


Московский государственный университет имени М. В. Ломоносова



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 Александр Геннадьевич*

# Содержание

<b>1</b>	<b>Введение</b>	<b>2</b>
<b>2</b>	<b>Методы отбора признаков</b>	<b>2</b>
2.1	Фильтры . . . . .	3
2.2	Встроенные алгоритмы . . . . .	3
2.3	Методы обертки . . . . .	3
<b>3</b>	<b>Алгоритмы отбора признаков с помощью Random Forest</b>	<b>4</b>
3.1	Boruta . . . . .	5
3.2	ACE . . . . .	6
<b>4</b>	<b>Пример использования алгоритмов отбора признаков</b>	<b>6</b>
4.1	Gene expression data . . . . .	6
<b>5</b>	<b>Эксперименты с данными</b>	<b>9</b>
5.1	Данные . . . . .	9
5.2	Оценка важности признаков . . . . .	10
5.3	Встроенный алгоритм . . . . .	10
5.4	Boruta . . . . .	12
<b>6</b>	<b>Заключение</b>	<b>13</b>

# 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, регрессии 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 (например, отбор генов, соответствующих определенному типу 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, то есть проверить все из возможных наборов признаков и выбрать те признаки, на которых ошибка минимальна. 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 отбора признаков - *фильтры*, *обертки* и *встроенные алгоритмы* [4].

## 2.1 Фильтры

*Фильтры (filters)* основаны на некоторых показателях, которые не зависят от метода классификации. Например такие как, корреляция признаков с целевым вектором, критерии информативности. Они применяются до классификации.

Одним из преимуществ фильтрации является то, что она может быть использована в качестве предварительной обработки для уменьшения размерности пространства и преодоления переобучения. Методы фильтрации, как правило, быстро работают.

Фильтры используются для отбора признаков в кластеризации, для построения начального приближения [2].

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 (embedded algorithms)* выполняют отбор признаков во время процедуры обучения классификатора, и именно они явно оптимизируют набор используемых признаков для достижения лучшей точности [10]. Преимущества 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 (wrappers)* 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 ис-

пользовать любой классификатор, который определяет степень важности признаков. Подробнее несколько алгоритмов обертки будут рассмотрены в этой работе далее.

### 3 Алгоритмы отбора признаков с помощью Random Forest

Random Forest — алгоритм машинного обучения, предложенный Leo Breiman и Adele Cutler. Представляет собой ансамбль многочисленных несмещенных, но чувствительных к обучающей выборке алгоритмов (деревьев решений). Каждый из этих классификаторов строится на случайном подмножестве объектов и случайном подмножестве признаков. Алгоритм сочетает в себе две основные идеи: метод бэггинга и метод случайных подпространств, предложенный Tin Kam Ho.

Пусть обучающая выборка состоит из  $N$  примеров, размерность пространства признаков равна  $M$ , и задан параметр  $m$ . Запишем пошагово алгоритм Random Forest.

Все деревья ансамбля строятся независимо друг от друга по следующей процедуре:

1. Сгенерируем случайную подвыборку с повторением размером  $n$  из обучающей выборки.
2. Построим решающее дерево, классифицирующее примеры данной подвыборки, причём в ходе создания очередного узла дерева будем выбирать признак, на основе которого производится разбиение, не из всех  $M$  признаков, а лишь из  $m$  случайно выбранных.
3. Дерево строится до полного исчерпания подвыборки и не подвергается процедуре прунинга (англ. pruning — отсечение ветвей)

Классификация объектов проводится путём голосования: каждое дерево ансамбля относит классифицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев.

Алгоритм Random Forest может быть использован в задаче оценки важности признаков. Для этого необходимо обучить алгоритм на выборке и во время построения модели для каждого элемента обучающей выборки посчитать out-of-bag-ошибку.

Пусть  $X_n^\ell$  - бутстрапированная выборка дерева  $b_n$ ,  $L(y, z)$  - функция потерь, тогда out-of-bag-ошибка вычисляется по следующей формуле:

$$\text{OOB} = \sum_{i=1}^{\ell} L \left( y_i, \frac{1}{\sum_{n=1}^N [x_i \notin X_n^\ell]} \sum_{n=1}^N [x_i \notin X_n^\ell] b_n(x_i) \right).$$

Затем для каждого объекта такая ошибка усредняется по всему случайному лесу. Чтобы оценить важность признака, его значения перемешиваются для всех объектов обучающей выборки и out-of-bag-ошибка считается снова. Важность признака оценивается путем усреднения по всем деревьям разности показателей out-of-bag-ошибок до и после перемешивания значений. При этом значения таких ошибок нормализуются на стандартное отклонение.

Кроме того, случайный лес имеет еще некоторые преимущества 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 каждый скрытый признак.
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 признаков 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 признаки 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 позволило сравнить результаты после использования Boruta.

Данные имеют информацию о 38 больных. По каждому из которых описано 3051 ген. Для оценки качества алгоритма было добавлено еще 1000 полусинтетических признаков, которые были сгенерированы 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

На Рис. 1 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.

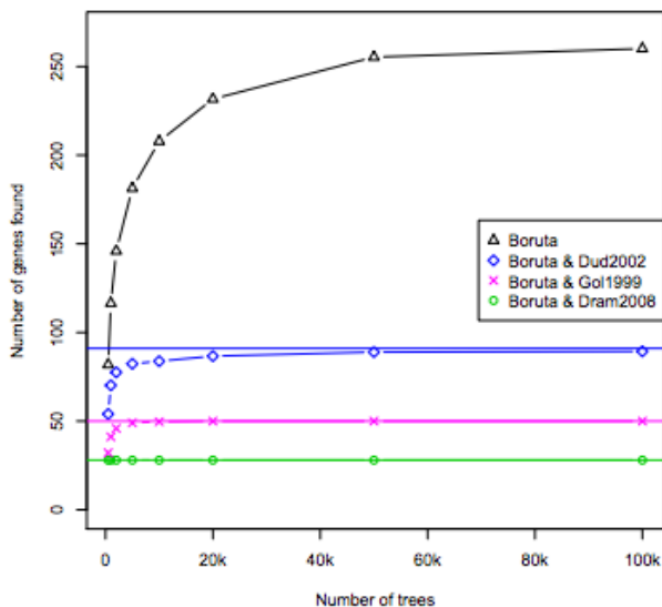


Рис. 1: зависимости количества выбранных генов, от числа деревьев



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 алгоритм Boruta эффективно справляется с задачей отбора признаков.

На Рис. 2 изображено, как пересекаются выбранные множества важных генов у различных алгоритмов используемых разными людьми в разное время.



Рис. 2: пересечение важных генов у различных алгоритмов

Видно, что множество генов выделенное с помощью Boruta с 500 деревьями пересекается со всеми выделенными генами с помощью результатов полученных ранее, а полученные ранее результаты слабо коррелировали между собой. Гены выделенные с помощью Boruta с 100000 деревьями включают в себя все гены выделенные остальными алгоритмами. Данный эксперимент 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* - Количество 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 доход рабочего;

*Average salary of employees in new businesses* - Средний доход 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 соудрудников в новых компаниях;

## 5.2 Оценка важности признаков

## 5.3 Встроенный алгоритм

Изначальная выборка была разбита на три части, для использования кросс-валидации по трем фолдам. На Рис. 3 приведен график в котором для каждого из 17 некатегориальных признаков показана их важность для трех частей выборки с помощью встроенного алгоритма в Random Forest

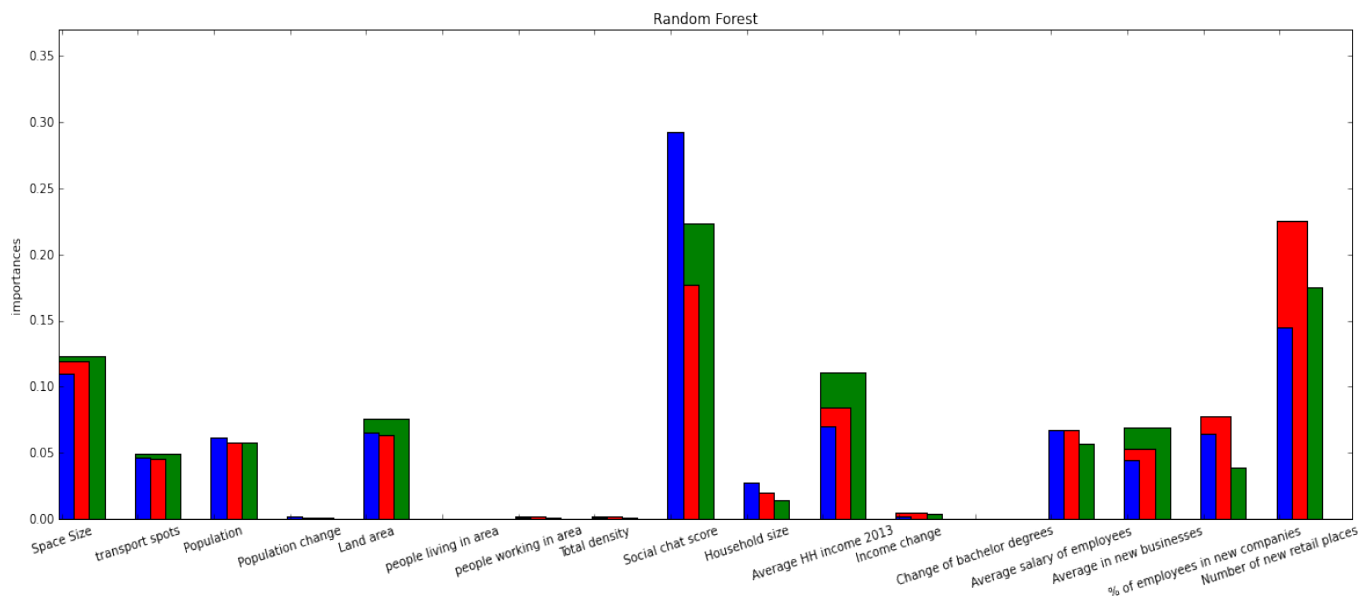


Рис. 3: важность признаков

На графике разными цветами указана важность признаков полученная по трем частям выборки. Максимальную важность в результате имеют следующие признаки:

*Social chat score* - Престижность ( социальный балл);

*Number of new retail places 2013 – 2010* - Число новых мест аренды;

*SpaceSize* - Размер помещения;

*Average HH income 2013* - Средний доход населения за 2013 год;

*Landarea* - Площадь региона;

Остальные признаки имели важность меньшую 0.05. Используя лишь те признаки, которые имели максимальную важность построен простой алгоритм, средняя ошибка которого по метрике MSE которого мало отличается от Random Forest. Данный алгоритм перебирает значение каждого важного признака у объекта и в зависимости от того в каком интервале лежат признаки присваивает целевой переменной определенную константу. Параметры настраивались на 2/3 случайно выбранных объектов из выборки, а тестировались на оставшейся 1/3.

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. На Рис. 4 приведены результаты тех алгоритмов, которые показали наиболее лучший результат.

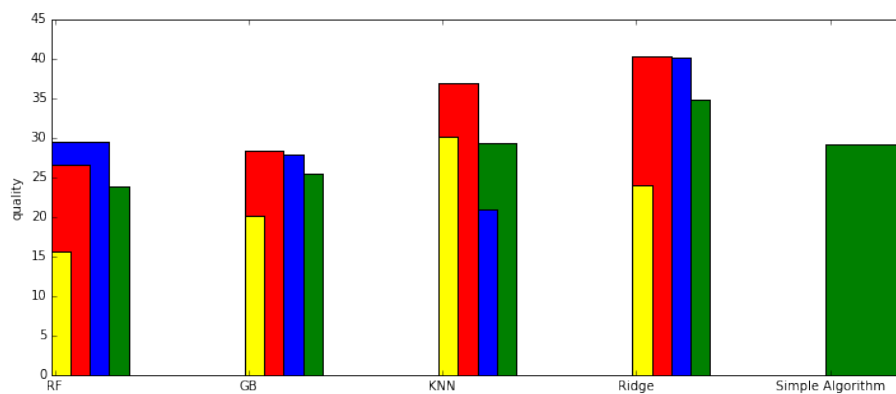


Рис. 4: результаты лучших алгоритмов

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

Реализация всех алгоритмов машинного обучения использовалась из библиотеки scikit-learn на языке python. Для каждого алгоритма были настроены все параметры с помощью функции GridSearchCV, но средняя ошибка по трем фолдам на настроенных параметрах почти не отличалась от ошибки на параметрах по умолчанию. Все алгоритмы запускались и на выделенных признаках с помощью описанных алгоритмов отбора признаков, но средняя ошибка по трем фолдам оставалась примерно такой же.

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.

## 5.4 Boruta

К данным был применен алгоритм Boruta. На Рис. 5 показана важность признаков, которую определил алгоритм Boruta. Зеленым цветом обозначены значимые признаки. Важность признаков определялась на 200 итерациях алгоритма.

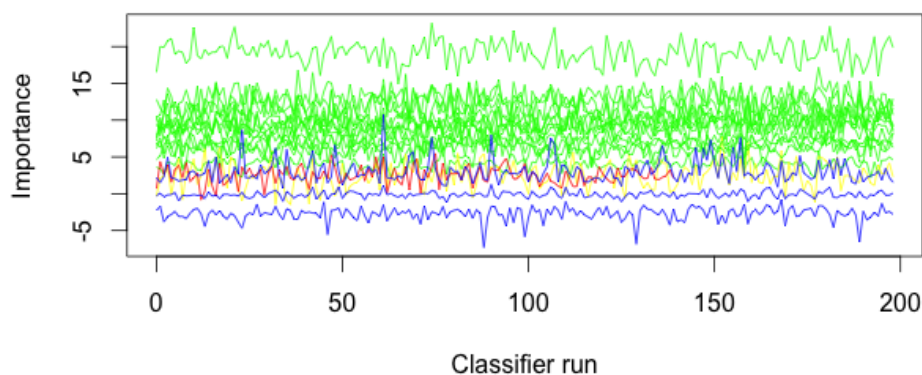


Рис. 5: важность признаков с помощью Boruta

Приведем наиболее значимые признаки, определенные с помощью Boruta.

<i>Признак</i>	<i>Значимость признака</i>
Space Size	18.6
Number of new retail places 2013-2010	12.8
Social chat score	12.3
Density of people working in area based on lat lon.	12.2
Total density living working.	11.3
Average HH income 2013	9.4
Household size	8.7
Population	8.2

Получилось что данный алгоритм выбрал почти все важные признаки, которые были выбраны с помощью встроенного алгоритма.

## 6 Заключение

Отбор признаков является важным этапом построения алгоритмов машинного обучения. Данный этап необходим, чтобы избавиться от шумовых признаков и благодаря этому улучшить качество и ускорить работу алгоритмов. Проведенные эксперименты 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)
- [11] Breiman, L.: Random forests. *Machine Learning* 45, 5–32 (2001)

- [12] L. Talavera. An evaluation of filter and wrapper methods for feature selection in categorical clustering. In In: proceeding of 6th International Symposium on Intelligent Data Analysis, pages 440–451, 2005. Madrid, Spain.