УДК 004.82

ОБЗОР СПОСОБОВ ПРИМЕНЕНИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ

Крюкова Я.Э.[1], Кручинин И.И.[1]

[1] Калужский филиал ФГБОУ ВО «Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)», Калуга, e-mail: fn1-kf@mail.ru

В статье был проведен обзор различных способов использования методов машинного обучения с целью прогнозирования поведения пользователей в сети. В терминах машинного обучения цель прогнозирования поведения пользователя состоит в том, чтобы предсказать исход Y из целого класса образцов поведения пользователя. Предсказание поведения пользователей подразумевает следующие требования для алгоритма прогнозирования: из-за высокой размерности пространства признаков необходим этап выбора функции, алгоритм должен рассчитать и предоставить прогноз достаточно быстро и, наконец, алгоритм должен иметь непрерывный характер обучения, с помощью которого он будет продолжать обучение в течение длительного времени. При прогнозировании поведения пользователей лучшие результаты представили такие методы как: логистическая регрессия, метод опорных векторов, метод ближайших соседей. Была дана краткая характеристика методам машинного обучения. Были рассмотрены примеры использования каждого из методов. С помощью логистической регрессии авторы работы [4] анализировали влияние особенностей вопросов на скорость откликов. С помощью метода опорных векторов авторы работы [5] изучали возможность более эффективного ранжирования сообщений в социальной сети Twitter. С помощью метода ближайших соседей авторы представили новый алгоритм совместной фильтрации на основе памяти для прогнозирования пользовательских отзывов о фильмах с использованием набора данных NetFlix. Кратко была рассмотрена каждая статья и результаты проделанной работы. В конце были сделаны соответствующие выводы.

Ключевые слова: прогнозирование поведения пользователей, логистическая регрессия, метод опорных векторов, метод ближайших соседей, машинное обучение

REVIEW OF THE METHODS OF APPLICATION OF MACHINE TRAINING METHODS FOR FORECASTING USER BEHAVIOR

Kryukova Ya.E.[1], Kruchinin I.I.[1]

[1] Bauman Moscow State Technical University (National Research University), Kaluga Branch, Kaluga, e-mail: fn1-kf@mail.ru

The article reviewed various ways of using machine learning methods to predict user behavior on the network. In terms of machine learning, the goal of predicting user behavior is to predict the outcome of Y from a whole class of user behavior patterns. Predicting user behavior implies the following requirements for the prediction algorithm: due to the high dimension of feature space, a function selection step is required, the algorithm must calculate and provide a prediction quickly enough and, finally, the algorithm must have a continuous learning nature, with which it will continue learning for a long time. When forecasting user behavior, the best results were presented by such methods as: logistic regression, support vector machine, nearest neighbor method. A brief description of machine learning methods was given. Examples of the use of each method were considered. Using logistic regression, the authors of work [4] analyzed the influence of question specifics on the response rate. Using the support vector method, the authors of [5] studied the possibility of more efficient ranking of messages on the social network Twitter. Using the closest neighbors method, the authors presented a new memory-based collaborative filtering algorithm for predicting custom movie reviews using the NetFlix dataset. We briefly reviewed each article and the results of the work done. At the end, appropriate conclusions were made.

Keywords: user behavior prediction, logistic regression, support vector machine, nearest neighbor method, machine learning

Введение. Машинное обучение является областью исследований, где изучаются методы построения алгоритмов для искусственного интеллекта. Подходы, основанные на машинном обучении, показали достойные результаты в прогнозировании поведения

пользователей в сети [1, 5, 9, 10]. Одним из направлений в создании моделей для прогнозирования поведения пользователя, является модель распространения информации в Интернете [11].

Действительно, изучая то, как данные распространяются в течение долгого времени, является полезным методом в понимании поведения пользователей в Интернете [2]. Другим направлением является использование различных функций данных для более детальной картины поведения пользователей. Первое направление дает лишь очень грубое представление о поведении пользователей в сети. Напротив, функциональный подход, основанный на модели на более низком уровне детализации, может обеспечить высокую точность модели [6, 12, 13].

Логистическая регрессия

Логистическая регрессия – это статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события путём подгонки данных к логистической кривой.

Основными преимуществами логистической регрессии являются: способность интерпретировать построенную модель, высокая производительность для больших наборов данных, а также многомерное пространство признаков.

В работе [4] авторы анализировали влияние особенностей вопросов на время откликов на сайте StackOverflow. Авторы использовали данные, собранные с сайта в течение четырех лет (с 2008 по 2012 года). Данные представлены на рисунке 1.

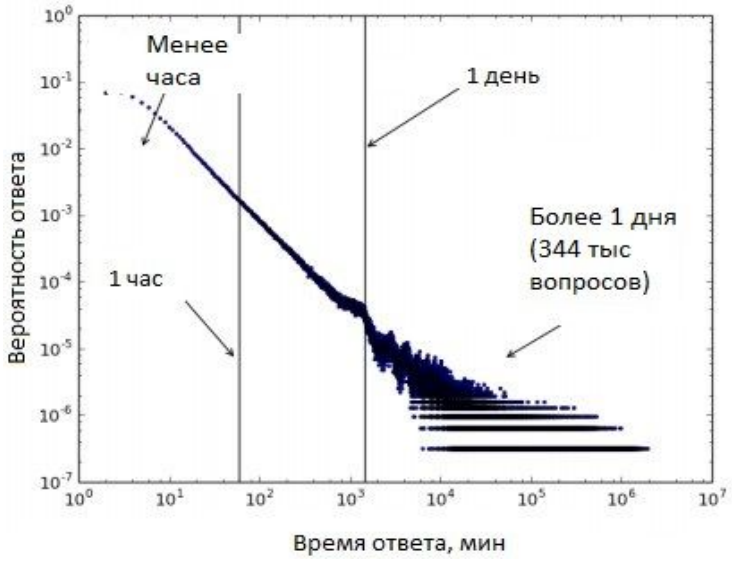

Рисунок 1. Зависимость вероятности ответа от времени

Авторы решали две задачи: спрогнозировать будет ли дан ответ на поставленный вопрос в течение 16 минут; спрогнозировать будет ли дан ответ на вопрос меньше чем за 1 час или будет ждать более 1 суток.

После обработки и анализа данных были определены факторы оказывающие наибольшее влияние на выходную функцию скорости и точности ответа: популярные тэги (столбец 2), процент активных подписчиков (столбец 3), конкретные (специфичные) тэги (столбец 4), все вышеперечисленные функции (столбец 5). Сравнение производилось с вопросом без тэга (столбец 1). Эксперимент проводился с помощью машинного обучения, основанного на таких методах как: логистическая регрессия, линейный и нелинейный метод опорных векторов и дерево решений. Результаты первого эксперимента представлены на рисунках Рисунок 2 и 3.

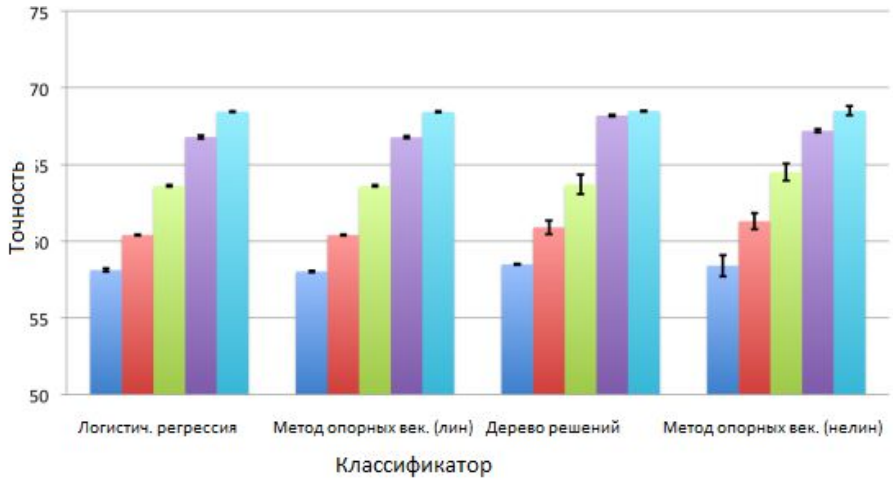

Рисунок 2. Точность ответа при решении первой задачи

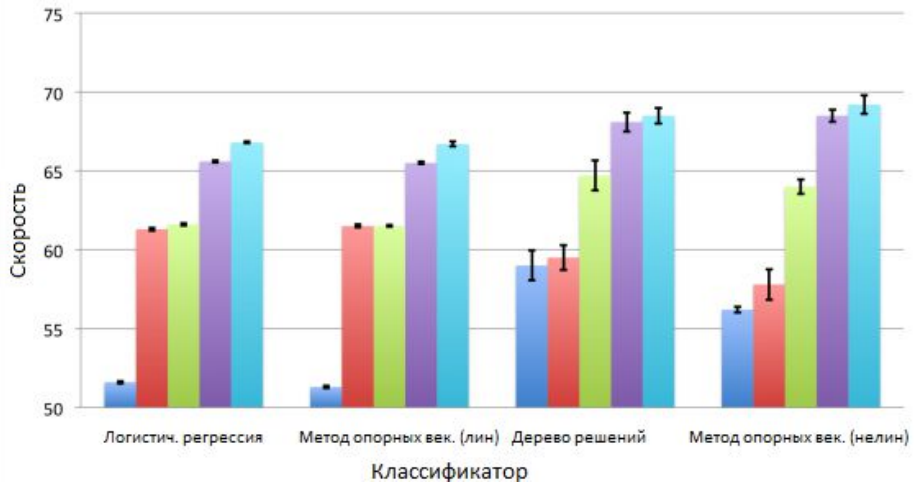

Рисунок 3. Скорость ответа при решении первой задачи

Результаты второго эксперимента – на рисунках 4 и 5.

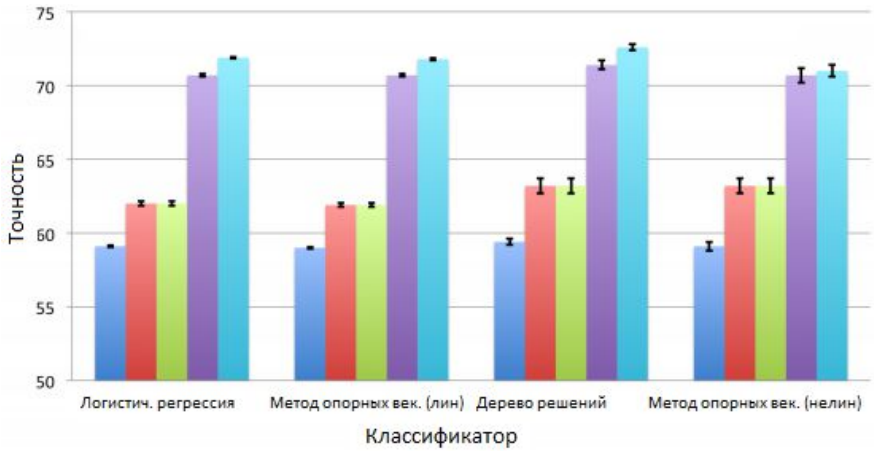

Рисунок 4. Точность ответа при решении второй задачи

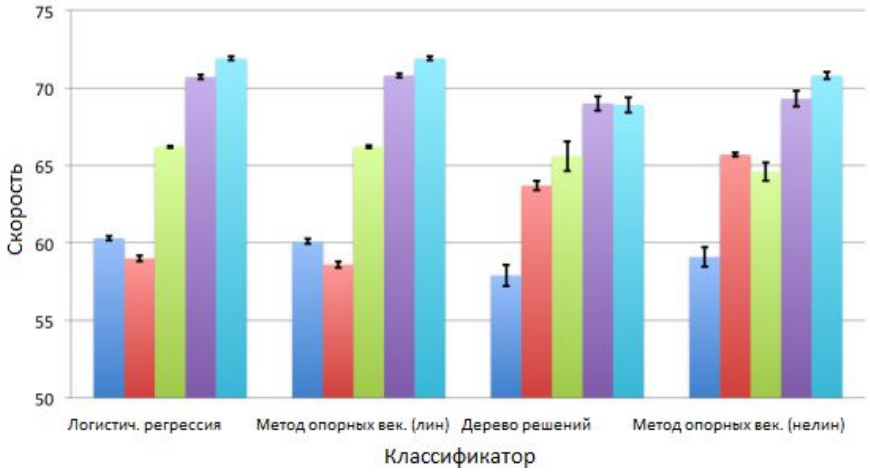

Рисунок 5. Скорость ответа при решении первой задачи

В результате решения обеих задач можно сделать одинаковые выводы: с ростом популярности тегов время отклика уменьшается, вопросы, содержащие слишком конкретные теги, остаются без ответа более длительное время, при увеличении количества подписчиков, время ответа уменьшается.

Метод опорных векторов

Метод опорных векторов – это набор алгоритмов, использующихся для задач классификации и регрессионного анализа. Учитывая, что в N-мерном пространстве каждый объект принадлежит одному из двух классов, данный метод генерирует (N-1) - мерную гиперплоскость с целью разделения этих точек на 2 группы.

В статье [7] авторы изучали проблему понимания важности сообщений в Twitter. Подход к данной проблеме состоял из двух этапов. Во-первых, была собрана статистика

об обмене сообщениями и ретвитами между пользователями за продолжительный промежуток времени (с 2006 по 2009 года). Была построена функция распределения (рисунок 6, рисунок 7) в зависимости от времени ожидания ответа и позиции сообщения.

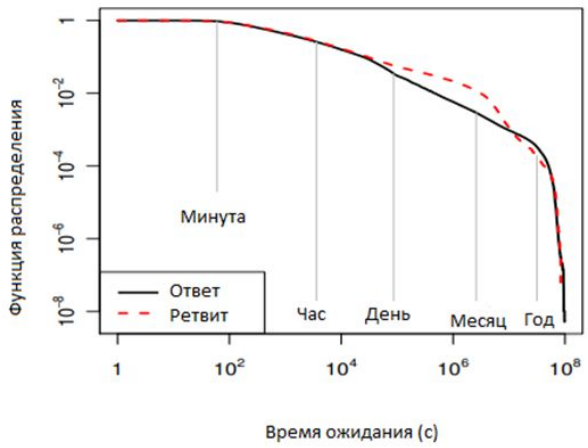

Рисунок 6. Функция распределения в зависимости от времени ожидания

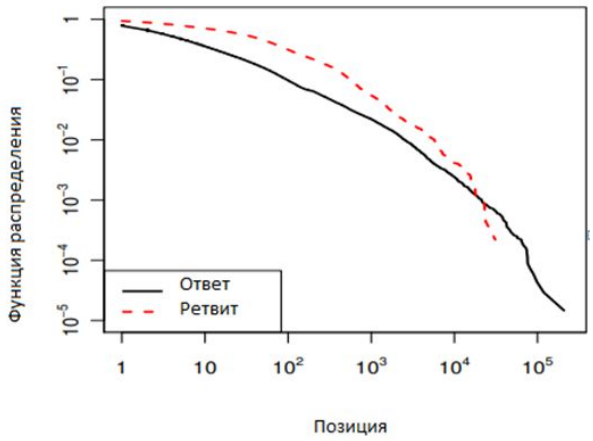

Рисунок 7. Функция распределения в зависимости от позиции сообщения

Затем определяли факторы, которые влияют на реакцию пользователя или вероятность ретвита. Наибольшее влияние на скорость ответа или ретрвита оказывают новизна сообщения, общались ли пользователи до этого или нет, скорость ответа отправителя, а также такие текстовые характеристики как: размер сообщения, наличие хэштегов, ретвитов, URL. Авторы показали, что некоторые из этих факторов могут быть использованы для улучшения представления порядка твитов в пользовательском интерфейсе.

Для подтверждения исходной гипотезы была проведена серия экспериментов, в котором сравнивались способы сортировки с помощью моделей, обученных с помощью

байесовского метода и метода опорных векторов с исходным способом сортировки. На рисунках 8 и 9 представлено сравнение количества ответов от позиций для различных способов сортировки для активных и пассивных пользователей

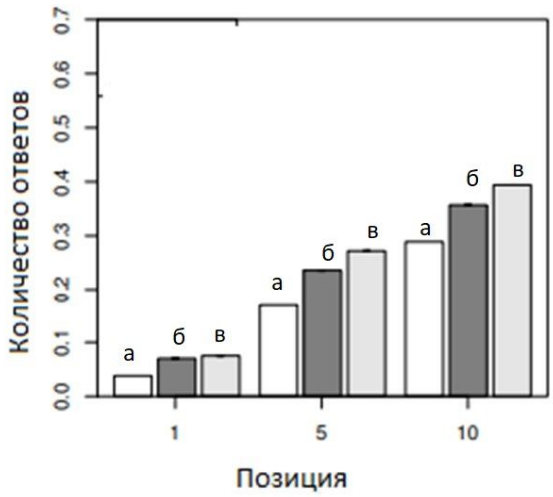

Рисунок 8. Количество ответов в зависимости от позиции для активных пользователей а) базовая сортировка; б) сортировка с помощью Баейсовского метода; в) сортировка с помощью метода опорных векторов.

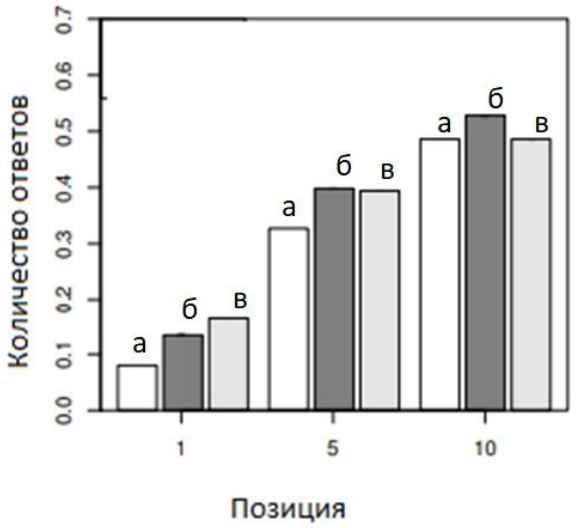

Рисунок 9. Количество ответов в зависимости от позиции для пассивных пользователей а) базовая сортировка; б) сортировка с помощью Баейсовского метода; в) сортировка с помощью метода опорных векторов.

Таким образом, из представленной диаграммы можно сделать следующие выводы: используя машинное обучение для изменения порядка сообщений можно достичь

большего количества ответов и ретвитов; сортировка с помощью метода опорных векторов показывает лучшие результаты для активных пользователей.

Метод ближайших соседей

Идея алгоритма метода ближайших соседей состоит в том, чтобы выбрать k – соседние векторы для входного вектора r как наиболее сходные с входным вектором.

В работе [8] авторы представили новый алгоритм совместной фильтрации на основе памяти для прогнозирования пользовательских отзывов о фильмах с использованием набора данных NetFlix.

Авторы получили список жанров из IMDB для примерно 5 миллионов фильмов, всего 23. Был выполнен простой поиск жанровых меток для всех фильмов NetFlix. Из-за различий в названиях фильмов и датах выпуска между базами данных авторы смогли добиться точного соответствия только для 6658 фильмов NetFlix. Далее был получен список из 33941 пользователей, которые оценили эти фильмы. Из них 3941 пользователь был отведен для целей тестирования, а остальные 30000 пользователей были использованы для определения оптимального значения параметра K в методе ближайших соседей с использованием 5-кратной перекрестной проверки. Каждый пользователь представлен в скрытом жанровом пространстве вектором признаков, накапливая вклад различных жанров фильмов, оцененных им. Для пользователей в наборе тестов один фильм на пользователя используется для прогнозирования метки, а остальные фильмы используются для генерации «частичного» вектора признаков. Тот же подход используется для проверки набора во время обучения. На рисунках 10 и 11 показан вектор признаков, созданный для двух пользователей с использованием их предпочтений фильма. Более высокий вес для таких жанров, как «Комедия», «Роман» и «Триллер» для пользователя 1, показывает симпатию к фильмам с этими конкретными темами, в то время как пользователь 2 предпочитает фильмы «Драма», «Криминал» и «Война».

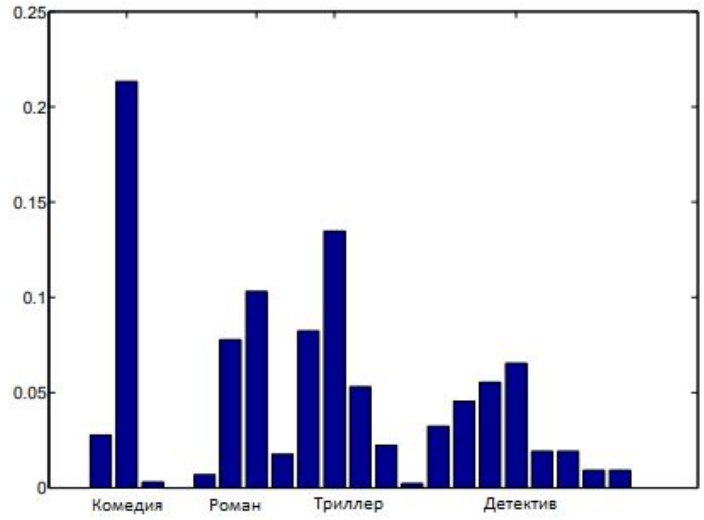

Рисунок 10. Результаты первого пользователя

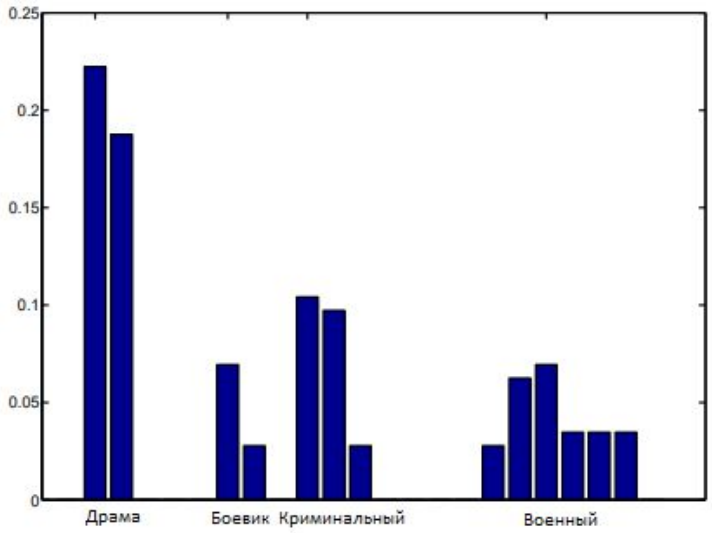

Рисунок 11. Результаты второго пользователя

Чтобы преодолеть проблему векторов функций голосования различной длины и уменьшить вычислительную сложность, работая с подмножеством «подходящих» пользователей, было введено пространство признаков, называемое пространством скрытого жанра. Для эффективного поиска в этом пространстве был использован метод ближайших соседей.

Для окончательной оценки рассмотрели два простых алгоритма для сравнения: первый быть средним числом всех голосов за фильмы (что отражает среднюю неспецифическую для пользователя популярность фильма) и наивным алгоритмом, который всегда рассматривает мнение трех рецензентов. На рисунке 12 показано сравнение ошибки данного алгоритма с двумя вышеперечисленными.

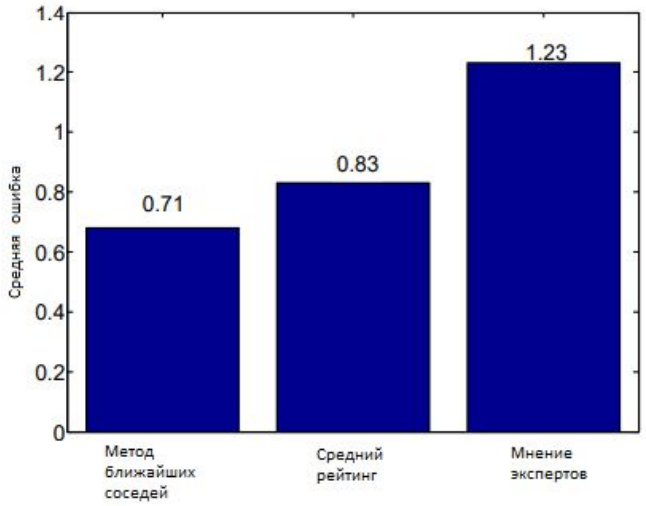

Рисунок 12. Сравнение алгоритмов, используя среднюю ошибку

Результаты данной работы являются многообещающими и превосходят два базовых метода, с которыми происходит сравнение. Недостатками данного алгоритма является, во-первых, жанр фильма может не содержать достаточной информации для решения данной задачи, а также модель полностью игнорирует пользовательские симпатии актерам и режиссерам.

Вывод. При выборе алгоритма, следует учитывать специфику анализируемых данных. Логистическая регрессия подходит для решения прогнозирующих задач благодаря масштабируемости и прозрачности интерпретации результатов.

Метод опорных векторов может показывать более высокую точность в прогнозировании, но он намного дороже в вычислительном плане и труден в интерпретации результатов прогнозирования.

Метод ближайших соседей прост в реализации, но неэффективно расходует вычислительные мощности вследствие необходимости хранения всей обучающей выборки и необходимости линейного сравнения классифицируемого объекта со всеми объектами выборки.

Список литературы

1. Аксютина Е.М., Белов Ю.С. Обзор архитектур и методов машинного обучения для анализа больших данных // Электронный журнал: наука, техника и образование. 2016. № 1 (5). С. 134-141.
2. Крюкова Я.Э., Гришунов С.С., Рыбкин С.В. Исследование современных моделей поведения пользователей в сети // Международный студенческий научный вестник. –

2019. – № 1. URL: http://www.eduherald.ru/ru/article/view?id=19538 (дата обращения: 10.03.2019).

3. Николенко С. И., Фишков А. А. SCM: новая вероятностная модель поведения пользователей интернет-поиска // Тр. СПИИРАН. № 20. 2012. С. 72–100.
4. Bhat V., Gokhale A. Analysis of question response time in Stack Overflow // International Conference on Advances in Social Networks Analysis and Mining. 2014, pp. 328–335.
5. Burlutskiy N., Petridis M., Fish A. Prediction of users' response time in q&a communities // International Conference on Machine Learning and Applications. 2015, pp. 618–623.
6. Choi S., Kim E., Oh S. Human behavior prediction for smart homes using deep learning // International Symposium on Robot and Human Interactive Communication. 2013, pp. 173–179.
7. Comarela G., Crovella M., Almeida V. Understanding factors that affect response rates in Twitter // In Proceedings of the 23rd ACM Conference on Hypertext and Social Media. 2012, pp. 123–132.
8. Goel D., Batra D. Predicting user preference for movies using Netflix database // Electrical and Computer Engineering, Carniege Mellon University, 2011, pp. 1–7.
9. Nazerfard E., Cook D. Using Bayesian networks for daily activity prediction // In AAAI Plan, Activity, and Intent Recognition Workshop. 2013, pp. 32–38.
10. Sadilek A., Krumm J. Far out: Predicting long-term human mobility // AAAI Conference on Artificial Intelligence. 2013, pp. 1–7.
11. Spiro E., Irvine C., DuBois C. Waiting for a retweet: modeling waiting times in information propagation. // Neural Information Processing Systems workshop of social networks and social media conference. 2012, pp. 1–8.
12. Zheng B., Thompson K., Lam S. Customers behavior prediction using artificial neural network // In Industrial and Systems Engineering Research Conference. 2013, pp. 700–709.
13. Zhu Y., Zhong E., Pan S. Predicting user activity level in social networks // In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. 2013, pp. 159–168.