Разработка и исследование алгоритмов определения предпочитаемых пользователем остановок общественного транспорта в геоинформационной системе на основе методов машинного обучения

А.А. Бородинов [1]

[1] *Самарский национальный исследовательский университет имени академика С.П. Королёва, 443086, Россия, г. Самара, Московское шоссе, д.34*

Аннотация

В работе рассматривается задача определения предпочитаемых пользователем остановок в рекомендательной транспортной системе. Проведено сравнение эффективности использования различных методов машинного обучения для решения указанной задачи в системе персонализированных рекомендаций: метода опорных векторов, дерева решений, случайного леса, AdaBoost, алгоритма k-ближайших соседей, многослойного персептрона. Сравнение указанных традиционных методов машинного обучения производилось также с предложенным методом, разработанным на основе алгоритма вычисления оценок. Экспериментальные исследования использовали реальные данные мобильного приложения «Прибывалка-63», являющегося частью сервиса tosamara.ru. Подтверждена как работоспособность, так и эффективность предложенного метода.

Ключевые слова: рекомендательная система, машинное обучение, пользовательские предпочтения.

Цитирование: **Бородинов, А.А.** Разработка и исследование алгоритмов определения предпочитаемых пользователем остановок общественного транспорта в геоинформационной системе на основе методов машинного обучения / А.А. Бородинов // Компьютерная оптика. – 2020. – Т. 44, № 4. – С. 646-652. – DOI: 10.18287/2412-6179-CO-713.

Citation: Borodinov AA. Development and research of algorithms for determining user preferred public transport stops in a geographic information system based on machine learning methods. Computer Optics 2020; 44(4): 646-652. DOI: 10.18287/2412-6179-CO-713.

Введение

Целью современных персонализированных рекомендательных систем является помощь пользователям с учётом их предпочтений и существующего пространственно-временного контекста. Одним из направлений в рекомендательных системах является навигация и транспорт. Соответствующие рекомендательные системы в общем случае должны предлагать пользователям подходящий им маршрут, прогнозировать потребности, пункт назначения, тип используемого транспорта и необходимые удобства во время поездки (кондиционирование воздуха, пандусы). С учётом предложенной информации пользователь и делает окончательный выбор. Этот выбор является основной информацией для обучения персонализированной рекомендательной системы. Учитывая значительную долю неопределенности при построении рекомендательных систем, неоднозначность выбора пользователя (может зависеть от настроения) и многих других усложняющих факторов, вопросы построения рекомендательных систем в настоящее время чрезвычайно актуальны и даже злободневны, если принять во внимание растущие объёмы поступающей информации.

В работе рассматривается проблема определения предпочитаемых пользователем остановок. С практической точки зрения перспективная персонализированная рекомендательная система, используя данные о пространственно-временных координатах пользователя (то есть времени и месте запроса), должна автоматически определить, какой остановкой общественного транспорта пользователь намерен воспользоваться в текущий момент для осуществления перемещения. Результатом работы такой системы может быть как ответ в виде конкретной остановки, так и ранжированный список остановок по убыванию «степени уверенности» их использования пользователем. Очевидно, что ответ «ближайшая остановка» в общем случае оказывается неудовлетворительным, поскольку выбор предпочитаемой остановки зависит от требуемой в данный момент корреспонденции пользователя.

При разработке рекомендательной системы существуют следующие основные проблемы:

- проблема холодного запуска – это хорошо известная проблема для рекомендательных систем [1]: важно достичь баланса между точностью рекомендуемых маршрутов при инициализации системы. Таким образом, приемлемое время настройки профиля личных предпочтений должно быть небольшим;
- недостаточно информации для анализа. Система располагает информацией только о пространственно-временных координатах. Необходимо учитывать информацию о состоянии транспортной инфраструктуры [2, 3], о погод-

ных условиях и о многом другом в разрабатываемых системах;
– многие факторы могут повлиять на выбор пользователя, например, погода или оборудование остановки. Даже если возможно решить проблему небольшого объёма информации, может возникнуть обратная проблема учёта большого количества факторов выбора в разрабатываемой системе и выбора подходящего метода машинного обучения.

В разрабатываемом модуле получения предпочтительной остановки решение проблемы холодного запуска системы представлено в параграфе 3.1. Решение двух других описанных проблем в данной работе не приводится в связи с отсутствием записи дополнительной информации в мобильном приложении «Прибывалка-63» и в сервисе tosamara.ru, однако представляет интерес для дальнейших исследований.

В настоящее время различные методы машинного обучения используются для создания рекомендательных систем. В данной работе представлены попытки использования методов машинного обучения для решения задачи выбора предпочтительной остановки общественного транспорта. Все исследования проводятся на основе данных, полученных с помощью мобильного приложения «Прибывалка-63» и сервиса tosamara.ru.

Работа построена следующим образом. В первом параграфе представлено современное состояние исследований в сфере рекомендательных систем. Подробное описание используемых данных, доступных с использованием приложения «Прибывалка-63», и рассмотренных методов машинного обучения, включая предложенный метод, представлено во втором параграфе. Эксперименты и их результат представлены в третьем параграфе. В заключении представлены оставшиеся проблемы, возможные решения и перспективные направления исследований.

1. Современное состояние исследований

В этом параграфе представлен краткий обзор подходов к построению рекомендательных систем с использованием различных алгоритмов машинного обучения. Хотя в некоторых статьях описывается построение системы рекомендаций для широкого круга задач, описанные методы подходят и для рассматриваемой задачи.

В статье [4] Ivens Portugal и соавторы приводят систематический обзор литературы, в которой анализируется использование алгоритмов машинного обучения в рекомендательных системах и определяются возможности для разработки программного обеспечения. Авторы рассмотрели 26 статей и определили, что в большинстве случаев используется байесовский подход или подход на основе дерева решений. Четырьмя наиболее изученными сферами применения рекомендательных систем являются фильмы, музыка, новостные и развлекательные статьи и товары на площадках онлайн-магазинов. В основном это связано с доступностью данных. Xiang Li и соавторы [5] предлагают многомерный контекстно-зависимый метод рекомендаций, основанный на улучшенном алгоритме случайного леса, который объединяет контекстную информацию для выполнения рекомендаций. В [6] Matthias Bogaert и соавторы оценивают методы многокритериальной классификации (классификации с множеством меток, англ. *multi-label classification*) в рекомендательных системах для перекрёстных продаж в секторе финансовых услуг. Авторы решили, что случайный лес является методом, который лучше остальных подходит для данной сферы рекомендательных систем. Haein Kim и соавторы [7] сравнивают дерево решений, нейронную сеть, метод опорных векторов, случайный лес и строят модель рекомендаций, отражающую последние модели покупок пользователей. Xibin Wang и соавторы [8] используют метод опорных векторов для определения предпочтения пользователя в фильмах на основе оценок, предоставленных им самим. Rachsuda Jiamthapthaksin и Than Htike Aung [9] предлагают методику профилирования предпочтений пользователя с использованием функций, созданных на основе поведения пользователя на странице Facebook и основанных на методе опорных векторов. В статье авторы также рассматривают наивный байесовский и нейросетевой подход. Mladen Marovic и соавторы [10] дают сравнение нескольких методов, основанных на алгоритме k-ближайших соседей и используемых в автоматическом прогнозировании рейтингов фильмов на основе данных IMDb. Igor Ivan и соавторы [11] статистически оценивают пешие расстояния до предпочтительной остановки общественного транспорта в двух выбранных региональных городах Чешской Республики. Кроме того, они изучили влияние демографических характеристик на расстояние, преодолеваемое пешком.

Исследования рекомендательных систем быстро развивались в течение последних нескольких лет, что подтверждает актуальность и необходимость дальнейших исследований. Однако область транспортных рекомендательных систем недостаточно изучена. Чтобы заполнить этот пробел в литературе, в настоящей работе проводится оценка общеизвестных методов машинного обучения и предложенного метода, основанного на алгоритме вычисления оценок, для определения предпочитаемых пользователем остановок общественного транспорта.

2. Методы определения предпочтительных остановок

В данном параграфе представлены возможные решения задачи определения предпочтительных для пользователя остановок общественного транспорта.

2.1. Данные запросов пользователей к мобильному приложению «Прибывалка-63»

В работе используются данные из мобильного приложения «Прибывалка-63». Данные в общем случае содержат следующую информацию:

– информация об остановках общественного транспорта (идентификаторы и координаты);
– информация о маршруте общественного транспорта (идентификаторы и список идентификаторов остановок);
– координаты пользователей и параметры запроса.

Каждый пользователь приложения выбирает остановку из списка и получает информацию о времени прибытия общественного транспорта к выбранной остановке. Такие информационные запросы показаны точками на карте. Остановка, к которой выполняется запрос, представлена звездой. Линии связывают соответствующие запросы и остановки. На рис. 1 представлена карта с запросами одного из пользователей к соответствующим остановкам.

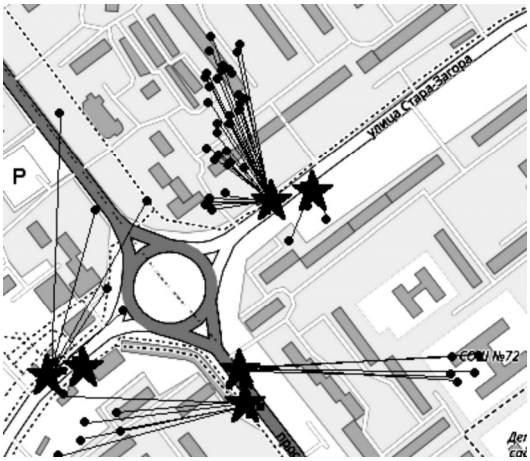

Рис. 1. Визуализация пользовательских запросов о прибытии общественного транспорта на остановки и соответствующих запросам остановок

Из указанной информации для нас представляет интерес та её часть, которая связана с конкретным пользователем и его запросами к остановкам. Каждый запрос конкретного пользователя (везде ниже пользователь предполагается фиксированным) можно представить в виде квартета:

$$\left(\mathbf{x}, d, t, ID(s)\right),$$

где $\mathbf{x}$ – координата местоположения пользователя, d – календарная дата запроса, t – время запроса, $ID(s)$ – идентификатор (условный номер) остановки, относительно которой произведён запрос пользователем. Вся информация о конкретном пользователе, собственно, представлена множеством указанных квартетов. Формально задача определения предпочтительной остановки заключается в определении / предсказании остановки (то есть идентификатора $ID(s)$) по пространственно-временному контексту, то есть по тройке $(\mathbf{x}, d, t)$. В более «развёрнутой» постановке по тройке $(\mathbf{x}, d, t)$ требуется получить ранжированный список идентификаторов остановок по убыванию некоторых «оценок» уверенности в использовании соответствующих остановок.

При интерпретации остановок как «классов» приходим формально к постановке задачи классификации или более широкой проблеме машинного обучения. Ниже рассмотрены предложенное решение указанной задачи и ряд решений с использованием хорошо известных методов машинного обучения. Дано их сравнение.

2.2. Предложенный метод, основанный на алгоритме вычисления оценок

Предлагаемый метод основывается на алгоритме вычисления оценок. Он применим как для решения задачи классификации, так и для построения ранжированного списка остановок общественного транспорта и был подробно описан автором настоящей работы в соавторстве в публикации [12]. Алгоритм вычисления оценок (кратко – АВО) был предложен академиком РАН Ю.И. Журавлёвым для решения задач распознавания в [13]. Алгоритм представляет собой мета-алгоритм / метод, в котором требуется определить набор функциональных и числовых параметров под конкретную задачу и критерий.

Определим значение оценки $\Gamma(features;\gamma)$, которое характеризует принадлежность вектора признаков к классу γ (в качестве классов в данном случае выступают идентификаторы остановок общественного транспорта), как

$$\Gamma\left(\mathbf{x}, d, t; \gamma\right) = \sum_{i \in \Im} \mu\left(\mathbf{x}, d, t; \mathbf{x}_i, d_i, t_i\right) I\left(\gamma_i = \gamma\right), \qquad (1)$$

где

$$\begin{aligned} &\mu\left(\mathbf{x}, d, t; \mathbf{x}_i, d_i, t_i\right) = \\ &= I\left(\left(w(d) \in W_0 \wedge w(d_i) \in W_0\right) \vee \right. \\ &\left. \vee \left(w(d) \in W_1 \wedge w(d_i) \in W_1\right)\right) \times \\ &\times \exp\left(-\alpha \left|t - t_i\right|\right) \cdot \exp\left(-\beta \left\|\mathbf{x} - \mathbf{x}_i\right\|\right). \end{aligned} \qquad (2)$$

Здесь $w(d) \in W$ – день недели, принимающий значения из множества:

$$\begin{aligned} &\mathrm{W} = \mathrm{W}_0 \cup \mathrm{W}_1, \\ &\mathrm{W}_0 \equiv \{MON, TUE, WEN, THU, FRI\}, \\ &\mathrm{W}_1 \equiv \{SAT, SUN\}, \end{aligned}$$

и индикатор события:

$$I(a) = \begin{cases} 1, & a = true; \\ 0, & a = false. \end{cases} \qquad (3)$$

Сначала вычисляем значения (1) для всех остановок из набора S:

$$\Gamma\left(\mathbf{x}, d, t; ID\left(s_i\right)\right), \quad i=\overline{1,|\mathrm{S}|}. \tag{4}$$

После этого полученный список сортируется путём уменьшения значений в (4), и получаем перестановку σ: $\mathbb{N}_{|S|} \to \mathbb{N}_{|S|}$. Полученная перестановка является решением задачи. Заметим, что данный метод позволяет получить упорядоченный список остановок, предоставляемый пользователю. Предложенный метод, по предварительно проведённым исследованиям, даёт наилучшие результаты на тестовой выборке при числовых параметрах $\alpha = 0{,}01$ и $\beta = 0{,}01$.

2.3. Репозиторий классических методов машинного обучения (классификаторов)

В качестве простейшего альтернативного метода решения был использован метод (*Mindistance*), основанный на выборе ближайшей остановки и, соответственно, упорядочивании остановок по возрастанию расстояния до них. Такой простой метод часто используется в существующих системах.

Метод опорных векторов (*SVM*) является одним из самых используемых методов среди рассматриваемых нами алгоритмов и чаще всего используется для сравнения. Функция разделения классов – это разделяющая гиперплоскость. Алгоритм максимизирует кратчайшее расстояние между точками, ближайшими к точкам на гиперплоскости. В этой статье используется радиальная базисная функция в качестве функции ядра. В работе используется реализация алгоритма из библиотеки scikit-learn [14] версии 0.22.2 на языке программирования Python 3.6.8.

Дерево решений (*Decision Tree*) – это структура иерархического типа, в которой ветви определяют раздел пространства признаков, а листы представляют собой элементарные функции классификации. Существуют различные методы построения деревьев. Коэффициент Джини используется в работе как критерий качества разбиения. Дерево создается до максимального размера без использования правила остановки, а затем оно обрезается. Алгоритм строит не одно, а последовательность вложенных усечённых деревьев. Затем выбирается наилучшее разбиение. В работе используется реализация алгоритма из библиотеки scikit-learn [14] версии 0.22.2 на языке программирования Python 3.6.8.

Случайный лес (*Random Forest*) – это набор деревьев решений. Решение о классификации принимается путём голосования большинством. Каждое дерево решений строится независимо. Для каждого дерева выбирается подмножество обучающего набора. Для расщепления дерева выбирается лучший атрибут. Как правило, дерево строится до исчерпания выборки, листья дерева должны содержать представителей только одного сорта. В работе используется реализация алгоритма из библиотеки scikit-learn [14] версии 0.22.2 на языке программирования Python 3.6.8.

Основная идея алгоритма *AdaBoost* состоит в том, чтобы обучить слабые классификаторы на наборе подготовленных данных. В роли слабых классификаторов могут выступать деревья решений. После этого прогнозы всех слабых классификаторов объединяются взвешенным большинством голосов для получения окончательного прогноза. На итерации повышения алгоритм изменяет веса для каждой выборки из обучающего набора. Первоначально весовые коэффициенты равны. Для каждой последующей итерации выборочные масштабы изменяются индивидуально. Вес данных, которые были классифицированы неправильно, увеличивается. Таким образом, алгоритм концентрирует обучение на труднодоступных данных. В работе используется реализация алгоритма из библиотеки scikit-learn [14] версии 0.22.2 на языке программирования Python 3.6.8.

Классификация *k-ближайших соседей* (*kNN*) является ещё одним часто используемым методом. Данный метод относит классифицируемый объект к тому классу, к которому принадлежат ближайшие k объектов выборки. В данной работе используется евклидова метрика. Несмотря на простоту подхода, метод хорошо проявил себя в сложных задачах классификации и регрессии. К недостаткам метода можно отнести необходимость хранения всех объектов обучающей выборки. Оптимальный выбор параметра k сильно зависит от количества и качества данных. В данной работе используется параметр $k = 1, 2, 3$. В работе используется реализация алгоритма из библиотеки scikit-learn [14] версии 0.22.2 на языке программирования Python 3.6.8.

Многослойный персептрон (*MLP*) приведён в качестве примера простой нейронной сети. MLP – контролируемый алгоритм обучения, который аппроксимирует функцию потерь. Этот подход отличается от логистической регрессии тем, что между входным и выходным слоями может быть один или несколько нелинейных слоёв, называемых скрытыми слоями. Каждый нейрон в скрытом слое преобразует значения из предыдущего слоя, используя взвешенное линейное суммирование, за которым следует нелинейная функция активации. В этой работе применена функция активации relu, на последнем слое – softmax. Нейронная сеть обучена с использованием алгоритма adam. В экспериментальных исследованиях использовалась архитектура нейронной сети с количеством скрытых слоёв от 1 до 3 и количеством нейронов на скрытых слоях от 10 до 100. Наилучший результат показала структура сети с 2 слоями по 100 нейронов на каждом слое. В качестве функции потерь применяется следующая функция:

$$Loss(\hat{y}, y, W) = -y \ln \hat{y} - (1-y)\ln(1-\hat{y}) + \gamma \left\| W \right\|_2^2,$$

где $\gamma \| W \|_2^2$ – L2 регуляризация. В работе используется реализация алгоритма из библиотеки scikit-learn

[14] версии 0.22.2 на языке программирования Python 3.6.8.

Описанные в данном подпараграфе методы, как правило, позволяют получить в результате решения только предпочтительную остановку, а не ранжированный список остановок (то есть решить только задачу классификации). Однако это позволяет осуществить сравнение методов, результаты которого представлены в следующем параграфе.

3. Экспериментальные исследования предложенного подхода

3.1. Предварительная обработка набора данных

Набор данных содержит информацию о времени запроса и координатах GPS мобильного устройства во время запроса. Далее временная метка была разделена на время дня в секундах и на день недели, что позволило определить выходные и рабочие дни. Запросы были записаны за четыре месяца. В мобильном приложении зарегистрировано 18441744 запроса от 116524 пользователей на 1479 остановок. Для экспериментов случайным образом были отобраны данные 300 пользователей с количеством запросов от 63 до 3083. Общее число запросов в используемых данных составляет 130262. Увеличение количества пользователей в используемом наборе данных значительно увеличило время проведения экспериментальных исследований и не повлияло на полученные результаты. Около половины отобранных пользователей имели среднее количество запросов около 160. Затем данные были разделены по запросам каждого пользователя на тренировочный набор и тестовый набор в соотношении 4:1. Чтобы получить достоверную оценку производительности модели, используется пятикратный метод перекрёстной проверки. Чтобы решить проблему системы холодного запуска, изначально пустой набор событий пополняется следующими значениями:

$$\left\{\left(\mathbf{x}_i, d0, t0; ID\left(s_i\right)\right)\right\} \cup \left\{\left(\mathbf{x}_i, d1, t0; ID\left(s_i\right)\right)\right\}, \quad i = \overline{1, |\mathrm{S}|}, \tag{5}$$

где $t0$ = «0час00мин», в качестве $d0$ и $d1$ выступают даты соответственно выходного и рабочего дней, предшествующих дате запуска «системы», а $\mathbf{x}_i (i = 1, |S|)$ – координаты остановок общественного транспорта $s_i (i = 1, |S|)$. Полученный обучающий набор холодного запуска был добавлен в обучающий набор после разделения полного набора данных на обучающую и тестовую выборки.

3.2. Исследование эффективности классификаторов при определении предпочтительных остановок

В экспериментальном исследовании была получена точность классификации для оценки методов машинного обучения. В табл. 1 представлены полученные результаты точности для каждого из классификаторов.

При проведении экспериментов после обучения на тренировочном наборе данных на вход исследуемым методам подавались пространственно-временные координаты пользователей из тестового набора данных. Результатом работы алгоритма являлась предпочитаемая пользователем остановка либо ранжированный список остановок для предложенного метода. Для расчёта точности бралась первая остановка из ранжированного списка. Точность отражает процент верно классифицированных предпочтительных остановок общественного транспорта. Истинными предпочтительными остановками считаются остановки из тестового набора данных.

Табл. 1. Результаты сравнения различных методов машинного обучения на тестовом наборе данных

Алгоритм	Точность
Предлагаемый метод	**64,223**
Ближайшая остановка	28,430
Метод опорных векторов	55,793
Дерево решений	54,399
Случайный лес	**69,825**
AdaBoost	39,908
kNN, k=1	63,155
kNN, k=2	62,860
kNN, k=3	**63,864**
Многослойный персептрон	43,483

Из результатов, приведённых в табл. 1, видно, что наилучшие показатели точности дают алгоритм, предложенный в подпараграфе 2.2, случайный лес и метод ближайшего соседа. Однако последние два метода позволяют получить только предпочтительную остановку общественного транспорта. А предложенный метод, основанный на алгоритме вычисления оценок, предоставляет упорядоченный список остановок, решая обе задачи.

Стоит отметить, что метод, основанный на определении ближайшей остановки, показал достаточно низкое значение точности по той причине, что пользователи чаще делают запросы к остановкам заранее, находясь, например, дома или на работе.

На рис. 2 представлена визуализация предпочитаемой остановки общественного транспорта для каждого запроса пользователя. Предпочитаемой остановкой считается первая остановка в упорядоченном списке, предоставляемом пользователю. При нахождении пользователя в каждой из заштрихованных областей предпочтительная остановка (первая остановка в ранжированном списке для методов, представленных в параграфе 2.2) будет одной и той же и соответствовать заштрихованной области. Для создания рисунка был использован метод определения предпочтительных остановок, основанный на алгоритме вычисления оценок.

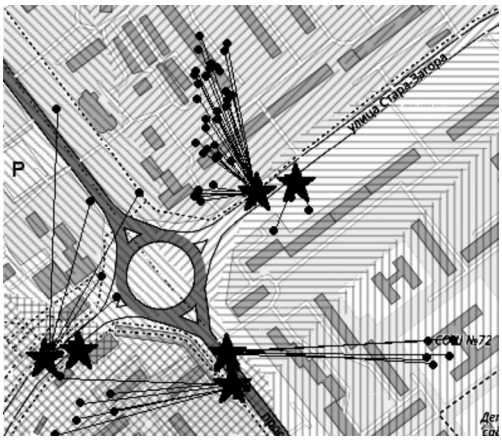

Рис. 2. Карта предпочтительных остановок, определённых предложенным методом, в зависимости от местоположения пользователя

Выводы и результаты

В статье представлено сравнение методов машинного обучения в задаче определения предпочтительной остановки пользователя, решаемой в рамках персонализированной рекомендательной системы. Получены показатели качества решения рассмотренной задачи для набора широко известных методов машинного обучения и для предложенного метода, основанного на алгоритме вычисления оценок. Результаты экспериментов показали, что наилучшими качественными показателями обладают предложенный метод и метод случайного леса. Однако предложенный метод, наряду с решением задачи классификации, то есть указания предпочтительной остановки, позволяет получить ранжированный список остановок. Указанная возможность делает его более предпочтительным, поскольку позволяет использовать его более гибко в персонализированной рекомендательной системе.

Перспективными направлениями исследований являются: использование информации о погоде, о состоянии транспортной сети и инфраструктуре, информации о сложности пути к остановке (светофоры, пандусы, пешеходные переходы) и другой дополнительной информации.

Благодарности

Работа финансировалась Министерством науки и высшего образования Российской Федерации (уникальный идентификатор проекта RFMEFI57518X0177).

Литература

1. **Campigotto, P.** Personalized and situation-aware multimodal route recommendations: The FAVOUR algorithm / P. Campigotto, C. Rudloff, M. Leodolter, D. Bauer // IEEE Transactions on Intelligent Transportation Systems. – 2017. – Vol. 18, Issue 1. – P. 92-102. – DOI: 10.1109/TITS.2016.2565643.
2. **Агафонов, А.А.** Исследование численного метода резервирования маршрутов в геоинформационной задаче маршрутизации автономных транспортных средств / А.А. Агафонов, В.В. Мясников // Компьютерная оптика. – 2018. – Т. 42, № 5. – С. 912-920. – DOI: 10.18287/2412-6179-2018-42-5-912-920.
3. **Агафонов, А.А.** Анализ больших данных в геоинформационной задаче краткосрочного прогнозирования параметров транспортного потока на базе метода k ближайших соседей / А.А. Агафонов, А.С. Юмаганов, В.В. Мясников // Компьютерная оптика. – 2018. – Т. 42, № 6. – С. 1101-1111. – DOI: 10.18287/2412-6179-2018-42-6-1101-1111.
4. **Portugal, I.** The use of machine learning algorithms in recommender systems: A systematic review / I. Portugal, P. Alencar, D. Cowan // Expert Systems with Applications. – 2018. – Vol. 97. – P. 205-227. – DOI: 10.1016/j.eswa.2017.12.020.
5. **Li, X.** A multi-dimensional context-aware recommendation approach based on improved random forest algorithm / X. Li, Z. Wang, L. Wang, R. Hu, Q. Zhu // IEEE Access. – 2018. – Vol. 6. – P. 45071-45085. – DOI: 10.1109/ACCESS.2018.2865436.
6. **Bogaert, M.** Evaluating multi-label classifiers and recommender systems in the financial service sector / M. Bogaert, J. Lootens, D. Van den Poel, M. Ballings // European Journal of Operational Research. – 2019. – Vol. 279, Issue 2. – P. 620-634.
7. **Kim, H.** An intelligent product recommendation model to reflect the recent purchasing patterns of customers / H. Kim, G. Yang, H. Jung, S.H. Lee, J.J. Ahn // Mobile Networks and Applications. – 2019. – Vol. 24, Issue 1. – P. 163-170. – DOI: 10.1007/s11036-017-0986-7.
8. **Wang, X.** Personalized recommendation system based on support vector machine and particle swarm optimization / X. Wang, J. Wen, F. Luo, W. Zhou, H. Ren. – In: KSEM 2015: Knowledge science, engineering and management / ed. by S. Zhang, M. Wirsing, Z. Zhang. – Cham, Heidelberg, New York, Dordrecht, London: Springer, 2015. – P. 489-495. – DOI: 10.1007/978-3-319-25159-2_44.
9. **Jiamthapthaksin, R.** User preferences profiling based on user behaviors on Facebook page categories / R. Jiamthapthaksin, T.H. Aung // 2017 9th International Conference on Knowledge and Smart Technology: Crunching Information of Everything, KST 2017. – 2017. – P. 248-253. – DOI: 10.1109/KST.2017.7886077.
10. **Marović, M.** Automatic movie ratings prediction using machine learning / M. Marović, M. Mihoković, M. Mikša, S. Pribil, A. Tus // MIPRO 2011 – 34th International Convention on Information and Communication Technology, Electronics and Microelectronics. – 2011. – P. 1640-1645.
11. **Ivan, I.** Factors influencing walking distance to the preferred public transport stop in selected urban centres of Czechia / I. Ivan, J. Horák, L. Zajíčková, J. Burian, D. Fojtík // GeoScape. – 2019. – Vol. 13, Issue 1. – P. 16-30. – DOI: 10.2478/geosc-2019-0002.
12. **Borodinov, A.A.** Analysis of the preferences of public transport passengers in the task of building a personalized recommender system / A.A. Borodinov, V.V. Myasnikov // CEUR Workshop Proceedings. – 2019. – Vol. 2391. – P. 198-205. – DOI: 10.18287/1613-0073-2019-2391-198-205.
13. **Журавлев, Ю.И.** Распознавание образов и распознавание изображений / Ю.И. Журавлев, И.Б. Гуревич. – В кн.: Распознавание, классификация, прогноз. Математические методы и их применение / под ред. Ю.И. Журавлева. – Вып. 2. – М.: Наука, 1989. – С. 5-72.
14. Supervised learning – scikit-learn 0.22.2 documentation [Electronical Resource]. – URL: https://scikit-learn.org/stable/supervised_learning.html (request date February 4, 2019).

Сведения об авторе

Бородинов Александр Александрович, в 2016 окончил Самарский национальный исследовательский университет им. академика С.П. Королева с отличием по специальности «Информационная безопасность автоматизированных систем». В настоящее время является аспирантом кафедры геоинформатики и информационной безопасности Самарского университета. Область научных интересов: обработка изображений, машинное обучение и распознавание образов, разработка рекомендательных систем. Страница в Интернете: https://ssau.ru/staff/304968395-borodinov-aleksandr-aleksandrovich/edu . E-mail: *aaborodinov@yandex.ru* .

ГРНТИ: 28.23.27, 28.23.35
Поступила в редакцию 2 марта 2020 г. Окончательный вариант – 7 мая 2020 г.

Development and research of algorithms for determining user preferred public transport stops in a geographic information system based on machine learning methods

A.A. Borodinov [1]

[1] *Samara National Research University, 443086, Samara, Russia, Moskovskoye Shosse 34*

Abstract

The paper considers a problem of determining the user preferred stops in a public transport recommender system. The effectiveness of using various machine learning methods to solve this problem in a system of personalized recommendations is compared, including a support vector method, a decision tree, a random forest, AdaBoost, a k-nearest neighbors algorithm, and a multilayer perceptron. The described traditional methods of machine learning are also compared with the method proposed herein and based on an estimate calculation algorithm. The efficiency and the effectiveness of the proposed method are confirmed in the work.

Keywords:. recommender system, machine learning, user preferences.

Citation: Borodinov AA. Development and research of algorithms for determining user preferred public transport stops in a geographic information system based on machine learning methods. Computer Optics 2020; 44(4): 646-652. DOI: 10.18287/2412-6179-CO-713.

Acknowledgements: The work was funded by the Ministry of Science and Higher Education of the Russian Federation (unique project identifier RFMEFI57518X0177).

References

[1] Campigotto P, Rudloff C, Leodolter M, Bauer D. Personalized and situation-aware multimodal route recommendations: The FAVOUR algorithm. IEEE Trans Intell Transp Syst 2017; 18: 92-102. DOI: 10.1109/TITS.2016.2565643.

[2] Agafonov AA, Myasnikov VV. Numerical route reservation method in the geoinformatic task of autonomous vehicle routing. Computer Optics 2018; 42(5): 912-920. DOI: 10.18287/2412-6179-2018-42-5-912-920.

[3] Agafonov AA, Yumaganov AS, Myasnikov VV. Big data analysis in a geoinformatic problem of short-term traffic flow forecasting based on a K nearest neighbors method. Computer Optics 2018; 42(6): 1101–1111. DOI: 10.18287/2412-6179-2018-42-6-1101-1111.

[4] Portugal I, Alencar P, Cowan D. The use of machine learning algorithms in recommender systems: A systematic review. Expert Syst Appl 2018; 97: 205-227. DOI: 10.1016/j.eswa.2017.12.020.

[5] Li X, Wang Z, Wang L, Hu R, Zhu Q. A multi-dimensional context-aware recommendation approach based on improved random forest algorithm. IEEE Access 2018; 6: 45071-45085. DOI: 10.1109/ACCESS.2018.2865436.

[6] Bogaert M, Lootens J, Van den Poel D, Ballings M. Evaluating multi-label classifiers and recommender systems in the financial service sector. Eur J Oper Res 2019; 279(2): 620-634. DOI: 10.1016/j.ejor.2019.05.037.

[7] Kim H, Yang G, Jung H, Lee SH, Ahn JJ. An intelligent product recommendation model to reflect the recent purchasing patterns of customers. Mob Netw Appl 2019; 24(1): 163-170. DOI: 10.1007/s11036-017-0986-7.

[8] Wang X, Wen J, Luo F, Zhou W, Ren H. Personalized recommendation system based on support vector machine and particle swarm optimization. In Book: Zhang S, Wirsing M, Zhang Z, eds. KSEM 2015: Knowledge science, engineering and management. 2015, p. 489–95. DOI: 10.1007/978-3-319-25159-2_44.

[9] Jiamthapthaksin R, Aung TH. User preferences profiling based on user behaviors on Facebook page categories, International Conference on Knowledge and Smart Technology: Crunching Information of Everything 2017: 248-253. DOI: 10.1109/KST.2017.7886077.

[10] Marović M, Mihoković M, Mikša M, Pribil S, Tus A. Automatic movie ratings prediction using machine learning. 34th International Convention on Information and Communication Technology, Electronics and Microelectronics 2011: 1640-1645.

[11] Ivan I, Horák J, Zajíčková L, Burian J, Fojtík D. Factors influencing walking distance to the preferred public transport stop in selected urban centres of Czechia. GeoScape 2019; 13(1): 16-30. DOI: 10.2478/geosc-2019-0002.

[12] Borodinov AA, Myasnikov VV. Analysis of the preferences of public transport passengers in the task of building a personalized recommender system. CEUR Workshop Proc 2019; 2391: 198-205. DOI: 10.18287/1613-0073-2019-2391-198-205

[13] Zhuravlev YuI, Gourevich IB. Pattern recognition and image recognition [In Russian]. In Book: Zhuravlev YuI, ed. Recognition, classification, prediction. Mathematical methods and their application. Issue 2. Moscow: "Nauka" Publisher; 1989: 5-72.

[14] Supervised learning – scikit-learn 0.22.2 documentation. Source: ⟨https://scikit-learn.org/stable/supervised_learning.html⟩.

Author's information

Aleksandr Aleksandrovich Borodinov, graduated with honours (2016) from Samara National Research University, majoring in Information Security of Automated Systems. He is currently a postgraduate of Geoinformatics and Information Security Department at Samara University. Research interests: image processing, machine learning and pattern recognition, recommendation systems. https://ssau.ru/staff/304968395-borodinov-aleksandr-aleksandrovich/edu.
E-mail: *aaborodinov@yandex.ru* .

Received March 2, 2020. The final version – May 7, 2020.