УДК 004.89

ПОДБОР ЗНАЧЕНИЙ ПАРАМЕТРОВ РАЗРАБОТАННОЙ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ, ОСНОВАННОЙ НА АНАЛИЗЕ ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ В СЕТИ

Крюкова Я.Э., Белов Ю.С.

Московский государственный технический университет имени Н.Э. Баумана, филиал, Калуга, e-mail: yasyok21@mail.ru

В настоящее время очень высока скорость развития электронной коммерции, социальных сетей и интернет-пространства в целом, что предоставляет пользователю огромное количество разрозненной информации. Имеющиеся самостоятельные средства систематизации и фильтрации не позволяют пользователям получить доступ к той информации, которую они ищут или которая заинтересовала бы их. Все выше описанное создало достаточно актуальную проблему поиска и отбора информации в связи с тем, что ориентироваться в постоянно растущих объемах данных все сложнее и сложнее. В результате чего получили свое распространение рекомендательные системы, то есть программы, которые стремятся предсказать, какие объекты: веб-сайты, новости, книги, фильмы и т.д. – интересны каждому конкретному пользователю, основываясь на предшествующем поведении. Цель этой статьи – отразить процесс подбора параметров разработанной модели. Проанализировать влияние выполняемых действий на результат работы системы в целом. И в заключение показать, насколько сильно изменяется качество рекомендаций одной и той же системы, что позволит сделать вывод о последствиях пренебрежения данными аспектом. Помимо этого, в статье выполнен анализ базы входных данных.

Ключевые слова: рекомендательные системы, анализ поведения пользователей в сети, подбор параметров нейронной сети, изменение ошибки обучения, изменение точности предсказания

SELECTION OF VALUES OF PARAMETERS OF THE DESIGNED RECOMMENDED SYSTEM BASED ON ANALYSIS OF THE BEHAVIOR OF USERS ON THE NETWORK

Kryukova Ya.E., Belov Yu.S.

Bauman Moscow State Technical University, branch, Kaluga, e-mail: yasyok21@mail.ru

Currently, the high speed of development of electronic commerce, social networks and the Internet space as a whole, which provides the user with a huge amount of disparate information. Available independent means of organizing and filtering do not allow users to access the information that they are looking for or that would interest them. All of the above has created an urgent problem of searching and selecting information, since it is more and more difficult to navigate in ever-growing volumes of data. As a result of this, recommender systems, that is, programs that seek to predict which objects are of interest to a particular user, based on previous user behavior, have gained distribution. The purpose of this article is to reflect the process of selecting the parameters of the developed model. To analyze the impact of the actions taken on the result of the system as a whole. In conclusion, to show how much the quality of recommendations of the same system changes, which will allow us to conclude about the consequences of neglecting this aspect. In addition, the article analyzes the input database.

Keywords: recommendation systems, analysis of user behavior on the network, selection of neural network parameters, change in learning error, change in prediction accuracy

Поведение пользователя достаточно абстрактная вещь, поэтому ни одна из метрик не сможет в полной мере описать мыслительные процедуры, происходящие в голове человека в процессе выбора. Но реализация работы рекомендательной системы пытается максимально точно проанализировать поведение человека, предшествующее отданию предпочтения тому или иному продукту.

Задача рекомендательной системы – проинформировать пользователя о товаре, который ему может быть наиболее интересен в данный момент времени. Клиент получает информацию, а сервис зарабатывает на предоставлении качественных требуемых услуг.

В рамках данного подхода рекомендации генерируются на основании интересов других похожих пользователей и непосредственно поведения человека, которому будет сделано предсказание.

Цель исследования: рассмотреть процесс подбора параметров рекомендательной системы. А затем проанализировать изменения качества и точности получаемых предсказаний.

Объект исследования: рекомендательные системы – это алгоритмы, которые способны распознавать такие закономерности, о существовании которых люди могут даже не подозревать.

Входные данные. В данной работе в качестве входных данных была использована база данных электронной торговли «events», ее часть представлена в таблице. Для сжатия входных данных и выделения в них зависимостей используется обработчик. Он позволяет облегчить процедуру обучения и улучшить качество прогнозирования [1–3].

Часть базы данных «events»

timestamp	visitorid	event	itemid	transactionid
1433221332117	257597	view	355908	
1433224214164	992329	view	248676	
1433221999827	111016	view	318965	
1433222276276	599528	transaction	356475	4000

Где в колонке timestamp отмечено время, а именно метка времени «юникс» – это вариант кодировки времени. Время было конвертировано таким образом для удобства работы с данным параметром. Так как юникс-время занимает меньше места, следовательно, легче обрабатывать базу в целом. К тому же его можно легко конвертировать в привычную дату.

Следующий столбец visitorid, или идентификатор пользователя – это уникальное значение каждого пользователя, которое присваивается ему в момент посещения сайта.

Далее идет столбец event, или событие – все действия пользователя можно разбить на три категории, а именно view – просмотр, addtocart – добавление в корзину и transaction – непосредственно покупка.

Столбец itemid, или идентификатор товара – это уникальный идентификатор каждого товара.

И заключительный столбец transactionid, или идентификатор покупки – это уникальное значение, присваиваемое каждой покупке.

Для наглядности рассмотрим последнюю строку таблицы. Она сообщает о том, что посетитель 599528 приобрел товар 356475, идентификатор данного события 4000, временная метка данного события 1433222276276.

Ошибка обучения. В качестве функции для вычисления ошибки обучения модели была выбрана функция кросс-энтропии, или логарифмическая функция потерь. Для вычисления ошибки обучения была выбрана данная функция, так как она позволяет измерить разницу между полученными и целевыми результатами [4; 5].

Если значение функции кросс-энтропии маленькое, следовательно, предсказание достаточно точное, если большое, то нет соответственно.

В случае большой величины функции ошибки необходимо ее уменьшить, для этого применяется инструмент для обновления весов нейронов optimizer с использованием метода Адам. И выполняется своеобразное обратное распространение ошибки обучения в сочетании с градиентным спуском с «импульсом» [6–8].

Предпочтение данному методу было отдано в связи с тем, что стандартный инструмент обратного распространения ошибки обучения имеет значительный недостаток при попадании в локальные минимумы, что приводит к ошибочным результатам в работе метода.

В основе метода Адам, как уже было сказано выше, заложена концепция «импульса или момента», за счет прибавления предыдущих градиентов к текущим, при слабом обновлении весов для типичных признаков и при небольшой скорости обучения [9; 10].

Результаты работы системы. Таким образом, после выполнения всех этапов обработки спроектированная система показывает результаты, представленные на рис. 1. Данный рисунок отражает предсказание, данное системой, то есть конкретные itemid, пользователю visitorid = 56. Для определения их достоверности было выполнено сравнение полученных предсказаний с целевыми данными.

Подбор параметров системы. Для подтверждения оптимальности архитектуры разработанной системы были проведены серии экспериментов, в которых изменялись ее параметры. В качестве параметров системы были выбраны следующие:

– количество нейронов в скрытом слое «hidden_dim»;

– количество слоев рекуррентной сети «number of RNN Layers»;

– количество эпох обучения «Number of epochs»;

– скорость обучения «Learning rate».

Параметрами первой модели были выбраны:

– количество нейронов в скрытом слое = 30;

– количество слоев рекуррентной сети = 1;

– количество эпох обучения = 50;

– скорость обучения = 0,05.

Результат представлен на рис. 2–4.

Параметрами второй модели были выбраны следующие:

– количество нейронов в скрытом слое = 15;

– количество слоев рекуррентной сети = 2;

– количество эпох обучения = 30;

– скорость обучения = 0,05.

Результат представлен на рис. 5–7.

```
In [98]: print(predict(model,Input_D[40:55], ID2Ord, Ord2ID)[0])

         [263513, 148666, 59635, 119736, 256706, 432171, 52039, 231796, 233439, 425920, 114763, 17478, 185847, 353436, 3313
         81]

In [101]: print(Target_D[40])

          56
          [263513 148666  59635 119736 256706 432171  52039 231796 233439 425920
           114763  17478 185847 353436 331381]
```

Рис. 1. Сравнение полученных предсказаний с целевыми данными

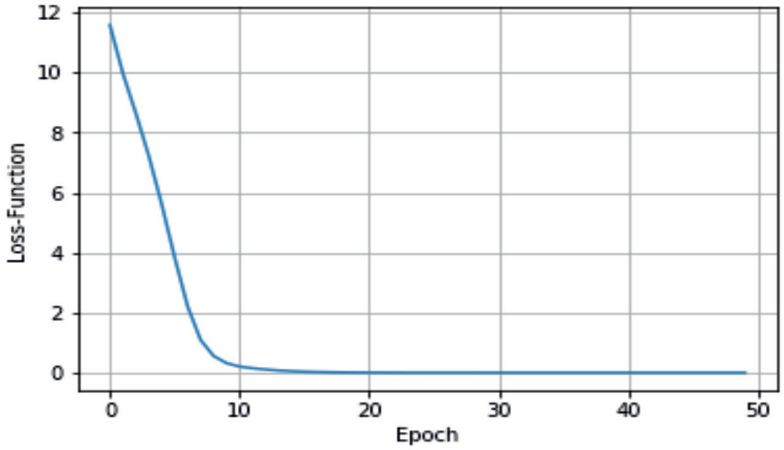

Рис. 2. Изменения ошибки обучения для первой модели

```
Epoch: 10/50.............. Loss: 0.3320 Accuracy: 93.3333%
Epoch: 20/50.............. Loss: 0.0208 Accuracy: 100.0000%
Epoch: 30/50.............. Loss: 0.0152 Accuracy: 100.0000%
Epoch: 40/50.............. Loss: 0.0152 Accuracy: 100.0000%
Epoch: 50/50.............. Loss: 0.0152 Accuracy: 100.0000%
```

Рис. 3. Значения изменения ошибки обучения и точности предсказания для первой модели

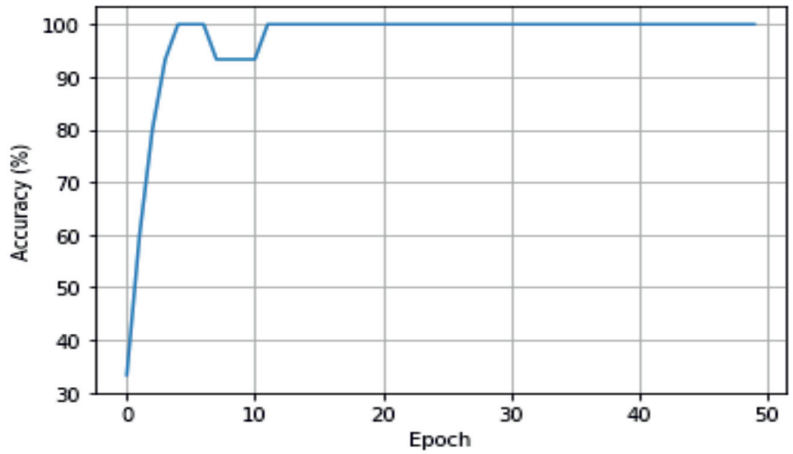

Рис. 4. График изменения точности предсказания для первой модели

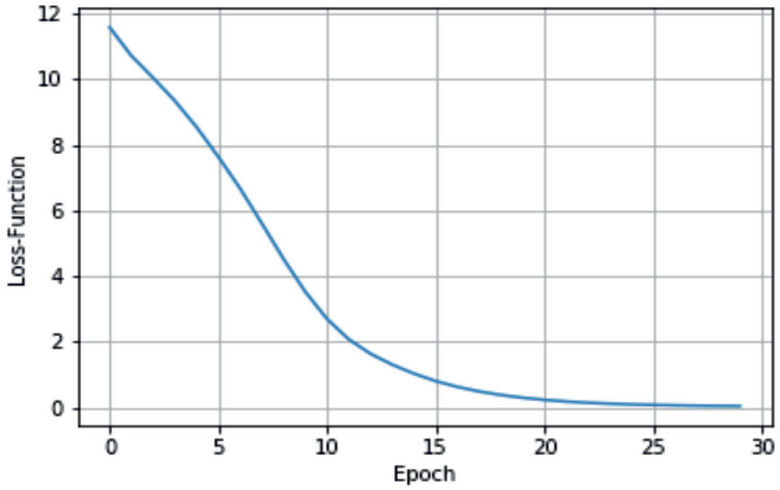

Рис. 5. Изменения функции ошибки обучения во второй модели

```
Epoch: 10/30.............. Loss: 3.5238 Accuracy: 73.3333%
Epoch: 20/30.............. Loss: 0.2920 Accuracy: 93.3333%
Epoch: 30/30.............. Loss: 0.0358 Accuracy: 80.0000%
```

Рис. 6. Значения изменения ошибки обучения и точности предсказания во второй модели

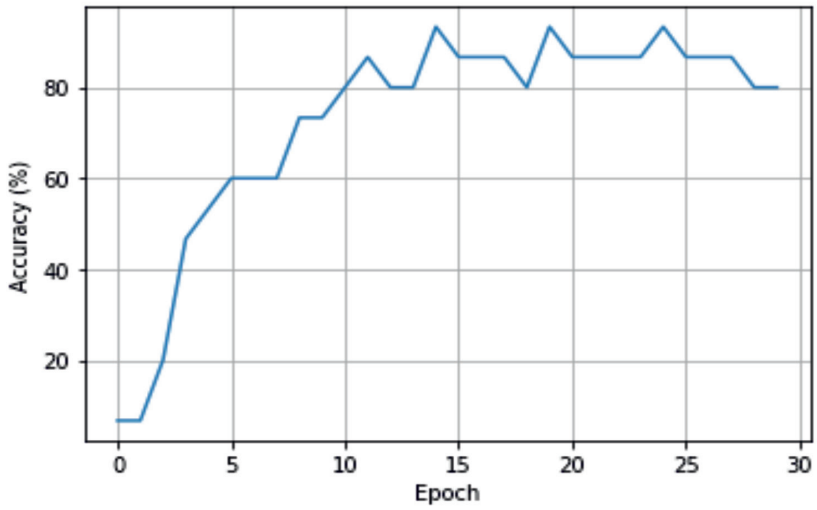

Рис. 7. Изменения точности предсказания во второй модели

Параметрами третьей модели были выбраны следующие:

– количество нейронов в скрытом слое = 15;
– количество слоев рекуррентной сети = 2;
– количество эпох обучения = 100;
– скорость обучения = 0,001.

Результат представлен на рис. 8–10.

Согласно полученным результатам первая архитектура является оптимальной. Что и отражают полученные на основании первой модели предсказания для пользователя с visitorid = 56.

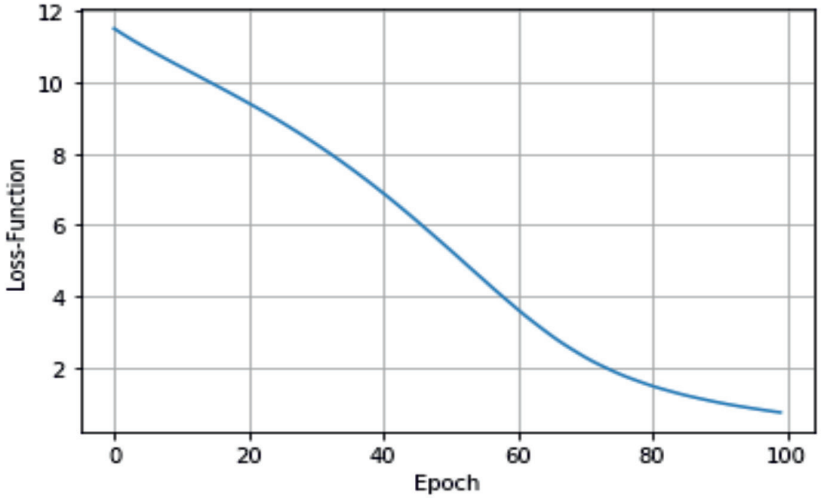

Рис. 8. Изменения ошибки обучения в третьей модели

```
Epoch: 10/100.............. Loss: 10.5242 Accuracy: 0.0000%
Epoch: 20/100.............. Loss: 9.5166 Accuracy: 20.0000%
Epoch: 30/100.............. Loss: 8.3891 Accuracy: 26.6667%
Epoch: 40/100.............. Loss: 7.0337 Accuracy: 46.6667%
Epoch: 50/100.............. Loss: 5.4462 Accuracy: 53.3333%
Epoch: 60/100.............. Loss: 3.7794 Accuracy: 73.3333%
Epoch: 70/100.............. Loss: 2.3918 Accuracy: 80.0000%
Epoch: 80/100.............. Loss: 1.5340 Accuracy: 86.6667%
Epoch: 90/100.............. Loss: 1.0402 Accuracy: 100.0000%
Epoch: 100/100............. Loss: 0.7315 Accuracy: 100.0000%
```

Рис. 9. Значения изменения ошибки обучения и точности предсказания в третьей модели

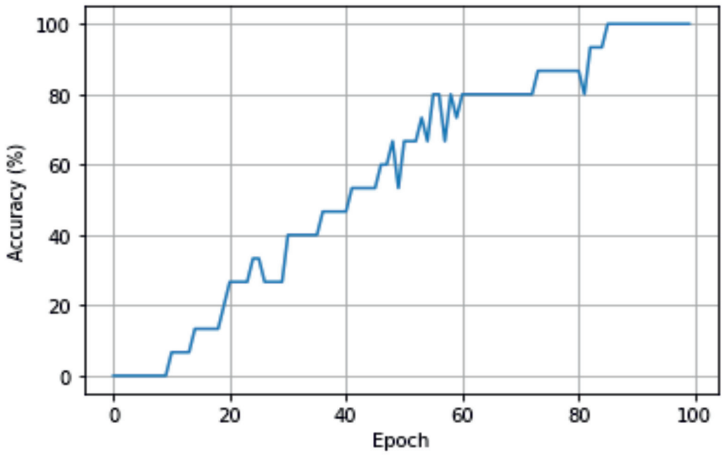

Рис. 10. Изменения точности предсказания в третьей модели

Заключение

По окончании обучения системы был произведен подбор параметров модели, для определения оптимальной архитектуры сети, основываясь на результатах, демонстрируемых полученными ошибками обучения и значениями точности предсказаний. Существует большое многообразие рекомендательных систем и алгоритмов их реа-

лизации. Но, несмотря на их пластичность, зачастую схожие системы дают значительно разнящиеся результаты, что отражается на работе системы в целом. В данном случае для разработанной системы была подобрана оптимальная архитектура для получения качественного предсказания.

Список литературы

1. Крюкова Я.Э., Гришунов С.С., Рыбкин С.В. Исследование современных моделей поведения пользователей в сети // Международный студенческий научный вестник. 2019. № 1. [Электронный ресурс]. URL: http://www.eduherald.ru/ru/article/view?id=19538 (дата обращения: 10.06.2020).

2. Крюкова Я.Э., Кручинин И.И. Обзор способов применения методов машинного обучения для прогнозирования поведения пользователей // Международный студенческий научный вестник. 2019. № 2. [Электронный ресурс]. URL: https://eduherald.ru/ru/article/view?id=19596 (дата обращения: 10.06.2020).

3. Крюкова Я.Э., Белов Ю.С. Прогнозирование поведения пользователей, основанное на машинном обучении // Актуальные вопросы науки: материалы юбилейной 50-й Международной научно-практической конференции (10.04.2019). М.: Изд-во «Спутник +», 2019. С. 165–167.

4. Николенко С.И., Фишков А.А. SCM: новая вероятностная модель поведения пользователей интернет-поиска // Тр. СПИИРАН. 2012. № 20. С. 72–100.

5. Николенко С.И., Фишков А.А. Обзор моделей поведения пользователей для задачи ранжирования результатов поиска // Тр. СПИИРАН. 2012. № 22. С. 139–175.

6. 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. New York, USA. 2012. P. 123–132.

7. Hu H., Liu B., Wang B., Exploring social features for answer quality prediction in CQA portals. In Machine Learning and Cybernetics International Conference on. Tianjin, China. 2013. P. 1904–1909.

8. Jernite Y., Halpern Y., Horng S., Predicting chief complaints at triage time in the emergency department. NIPS Workshop on Machine Learning for Clinical Data Analysis and Healthcare. New York, USA. 2013. P. 1–5.

9. Raikwal J., Singhai R., Saxena K., Article: Integrating markov model with knn classification for web page prediction. International Journal of Computer Applications. Berlin, Heidelberg. 2013. P. 313–323.

10. Scellato S., Noulas A., Mascolo C., Exploiting place features in link prediction on location-based social networks. In Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining. New York, USA. 2011. P. 1046–1054.