ПРИМЕНЕНИЕ КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ ДЛЯ ВЫЯВЛЕНИЯ ОШИБОК ПЕРСОНАЛА ПРИ ВВОДЕ ДАННЫХ ДЛЯ ОЦЕНКИ ФУНКЦИОНАЛЬНОГО РЕСУРСА ОБЪЕКТОВ ТРАНСПОРТНОЙ ИНФРАСТРУКТУРЫ ОАО «РЖД»

Моисеенко Илья Владимирович

инженер-электроник, ГУП «Московский метрополитен», РФ, г. Москва

Менакер Константин Владимирович

канд. техн. наук, доцент кафедры «Электроснабжение» Иркутского государственного университета путей сообщения, РФ, г. Иркутск

Орлов Александр Валерьевич

канд. техн. наук, доцент кафедры «Системы управления транспортной инфраструктурой» Российского университета транспорта (МИИТ), РФ, г. Москва

Орлов Виктор Валерьевич

директор направления «Безналичные решения» Байкальский банк ПАО «Сбербанк», РФ, г. Иркутск

USE OF COLLABORATIVE FILTERING TO IDENTIFY HUMAN ERRORS IN DATA ENTRY TO ESTIMATE THE FUNCTIONAL RESOURCE OF RUSSIAN RAILWAYS TRANSPORT INFRASTRUCTURE FACILITIES

Ilya Moiseenko

State Unitary Enterprise "Moscow Metro" electronic engineer, Russia, Moscow

Konstantin Menaker

Ph.D. in Engineering Science, senior lecturer of the department "Electricity Supply" of Irkutsk State University of Railways, Russia, Irkutsk

Alexander Orlov

Ph.D. in Engineering Science, senior lecturer of the department «Transport infrastructure management systems» of the Russian University of transport (MIIT), Russia, Moscow

Viktor Orlov

Director of Cashless Solutions, Baikal Bank Sberbank PJSC, Russia, Irkutsk

Аннотация. Ошибки персонала при вводе данных значительно влияют на точность оценки функционального ресурса объектов транспортной инфраструктуры ОАО «РЖД». Для контроля ввода и указания персоналу на возможные ошибки рассматривается возможность применения

модели коллаборативной фильтрации. Модель основана на гипотезе о наличии эталонных схем заполнения данных, которые выявляются на основе формируемых в результате статистической обработки эталонных данных критериев сходства и отбора и используются для прогноза значений при вводе новых данных.

Abstract. Human errors during data entry significantly affect the accuracy of the assessment of the functional resource of the transport infrastructure facilities of Russian Railways. To monitor input and to indicate possible errors to personnel, the possibility of using a collaborative filtering model is considered. The model is based on the hypothesis of the presence of reference data filling schemes, which are detected on the basis of the criteria of similarity and selection formed as a result of statistical processing of reference data and are used to predict values when entering new data.

Ключевые слова: коллаборативная фильтрация; выявление ошибок; информационные системы; big data; объекты транспортной инфраструктуры.

Keywords: collaborative filtering; error detection; information systems; big data; transport infrastructure.

В настоящее время в компании ОАО «РЖД» в рамках цифровой трансформации активно развивается методология управления ресурсами, рисками и анализами, и надежностью (УРРАН). Она призвана обеспечить повышение эффективности технической эксплуатации объектов транспортной инфраструктуры (ОТИ) и подвижного состава. В основе используемых методов, расчетов и принятия решений лежит обработка больших объемов данных, формально относящихся к категории Big Data (большие данные).

Несмотря на наличие большого количества средств автоматической регистрации, ввод значительной доли первичных данных все еще осуществляет персонал, кроме того в информационных системах уже используется большое количество ранее внесенных персоналом данных. Глубокая проверка объективности вводимых в информационные системы данных вплоть до настоящего времени практически не осуществляется, поэтому количество ошибочных данных велико.

Важной задачей, решаемой в рамках методологии УРРАН, где проблема ручного ввода исходных данных оказывает большое влияние на результат, является оценка функционального ресурса ФР. ФР характеризует объект в части количества и качества реализуемых функций в потенциальных условиях эксплуатации. Результат оценки ФР учитывается при принятии руководством относящихся к технической эксплуатации решений по назначению капитальных ремонтов, модернизации, замене или продлению эксплуатации ОТИ и влияет на производственное планирование.

Для оценки ФР у ОТИ персоналу требуется вручную заполнить ряд форм, указав какие из требуемых функций ОТИ могут быть реализованы, а какие – нет. Всего одна ошибка может полностью изменить оценку ФР у ОТИ, тогда как количество функций у ОТИ весьма велико. Например, для ОТИ относящихся к железнодорожной автоматике и телемеханике (ЖАТ) требуется указать более 50 функций, причем, для каждого ОТИ на станции и перегоне – отдельно. Количество ОТИ ЖАТ на сети превышает 10 тысяч. Поэтому ошибки неизбежны, особенно в условиях дефицита времени у персонала и их требуется выявлять.

Анализ заполненных форм для оценки ФР показал, что в них можно выделить ряд эталонных схем заполнения, различных для разных типов ОТИ, а также условий эксплуатации, обусловленных классом и специализацией железнодорожной линии. При этом между отдельными формами могут сохраняться локальные отличия, а сами формы могут добавляться при появлении новых ОТИ.

В связи с этим, целесообразно применить коллаборативную фильтрацию – модель,

относящуюся к искусственному интеллекту и имеющую возможность обучения и адаптации. При коллаборативной фильтрации прогнозное значение для некоторой позиции в новой (текущей) форме определяется на основе значений на этой же позиции в ранее заполненных формах с учетом меры их сходства с новой.

Модель реализуется в два этапа. На первом этапе квалифицированным персоналом осуществляется отбор, проверка на валидность и ввод в модель эталонных исходных данных о значениях, принимаемых каждой из функций, используемых для оценки ФР. На втором этапе для каждого нового набора данных модель формирует критерии сходства и отбора, на основе которых прогнозируется очередное значение функции.

Каждая функция, учитываемая при оценке ФР, может иметь только два значения: «да» и «нет». Ее значения требуется закодировать: «да» - 1, «нет» - 0. Вместо названий функций целесообразно указать их порядковые номера при их сквозной нумерации, а ОТИ допустимо обозначить буквами.

Фрагмент таблицы исходных данных представлен на рисунке 1.

№ функций	Эталонные ОТИ						Текущий ОТИ
	А	Б	В	Г	Д	Е	у
1	1	1	1	0	1	0	1
2	1	1	1	1	1	1	1
…	…	…	…	…	…	…	…
n	0	0	1	0	0	1	?

Рисунок 1. Фрагмент таблицы с исходными данными

Перед использованием модели исходные данные требуется подвергнуть стандартной процедуре нормализации.

Для вычисления прогнозного значения некоторой функции у нового набора данных (значок «?» на рисунке 1), соответствующего условному «текущему» объекту требуется сформировать критерии сходства и отбора. В общем случае критерии сходства и отбора могут быть основаны на различных мерах: косинусной мере, коэффициенте корреляции Пирсона, коэффициенте Танимото и ряде других.

Для оценки ФР в качестве меры сходства применен коэффициент корреляции Пирсона между y новым набором данных, образованных столбцом данных текущего объекта, и i-ым x_i эталонным набором:

$$r_{x_i y} = \frac{cov(x_i, y)}{\sqrt{s^2(x_i) \cdot s^2(y)}}, \quad (1)$$

$cov(x_i, y)$ x_i y

где - ковариация между наборами и ;

$s\ (x_i), s\ (y)$

- стандартные отклонения наборов соответственно.

Прогнозное значение *j*-ой функции ОТИ представляет собой округленный до ближайшего целого результат расчета по формуле:

$$y_j = \frac{\sum_{i=1}^{x} x_{ij} \cdot (r_{x_i y} + 1)}{\sum_{i=1}^{x} (r_{x_i y} + 1)}$$

. (2)

Критерий отбора формулируется следующим образом: в расчете участвуют только

x $r_{x_i y} \geq Z$ Z

те эталонных наборов данных, для которых: (число из диапазона от 0,3 до 0,9, задающее допустимое сходство)

Результат оценки схожести данных текущего ОТИ с эталонными (см. рисунок 1), выполненный по формуле 1, приведен на рисунке 2.

Схожесть набора данных текущего ОТИ с эталонными наборами					
с А	с Б	с В	с Г	с Д	с Е
0,36	-0,04	0,42	0,24	0,36	0,07

Рисунок 2. Схожесть текущего с каждым из эталонов

Как видно, наибольшая схожесть у текущего набора имеется с наборами А, В и Г. Их значения в соответствии с формулой 2 войдут в прогноз для текущего ОТИ с наибольшими весами. Так для строки *n* (см. рисунок 1) значение функции у ОТИ А и Д составило 0, а В – один. Средневзвешенное по формуле 2 после округления формирует прогнозный результат для обозначенного вопросом значения функции текущего ОТИ – 0.

Это значение может быть использовано следующим образом:

1. Если в данном месте имеется пропуск, то его следует заполнить данным результатом.

2. Если персонал пытается ввести на данной позиции 1, то предварительно показать ожидаемое число 0, а после ввода выделить единицу, например, цветом.

Опыт практического использования коллаборативной фильтрации для оценки ФР показал, что при вводе всем функциям требуется задать ранги, позволяющие определить порядок их предъявления для заполнения персоналом с целью идентификации нового набора.

В качестве меры, формирующей ранг можно использовать разные функции: выборочную дисперсию, количество информации, энтропию и проч.

При оценке ФР для формирования рангов авторами использовалась выборочная дисперсия:

самые высокие ранги присваивались функции, с наибольшей дисперсией среди эталонных наборов данных, самые маленькие – с наименьшей.

Применение коллаборативной фильтрации показало свою пригодность для решения задачи контроля вводимых персоналом данных для оценки ФР. Данная модель в таком же виде или модифицированном может быть использована и для выявления ошибок в иных вводимых данных для нужд методологии УРРАН и других задач, но после проведения соответствующих исследований.

Список литературы:

1. Князева А.А Способы построения гибридной рекомендательной системы на основе данных о заказах библиотеки // Князева А.А., Колобов О.С., Турчановский И.Ю / Труды XVII Международной конференции DICR-2019, Новосибирск, 3-6 декабря 2019 г. 96-101 С.