http://dx.doi.org/10.35596/1729-7648-2022-20-2-46-52

Оригинальная статья / Original paper

УДК 004.852

ВЛОЖЕННОЕ ПРЕОБРАЗОВАНИЕ С СОХРАНЕНИЕМ СЕМАНТИКИ ИСХОДНЫХ ДАННЫХ

М.Е. ВАТКИН., Д.А. ВОРОБЕЙ., М.В. ЯКОВЛЕВ., М.Г. КРИВОВА

ОАО «Сбер Банк», г. Минск, Республика Беларусь

Поступила в редакцию 27 октября 2021

Аннотация. В современном мире данные, используемые для описания объектов, часто представлены в виде разряженных векторов с большим количеством признаков. Работа с такими данными является вычислительно неэффективной, что зачастую приводит к переобучению при моделировании. Поэтому используются алгоритмы понижения размерности данных, одними из которых являются автокодировщики. В статье предложен новый подход для оценки свойств полученных векторов меньшей размерности, а также основанная на этом подходе функция потерь. Идея предложенной функции потерь состоит в вычислении качества сохранения семантической структуры в пространстве вложений и добавлении этой метрики в функцию потерь, что позволяет сохранить отношения объектов в пространстве вложений и таким образом сохранить больше полезной информации об объектах. Полученные результаты показывают, что использование комбинации среднеквадратичной функции потерь вместе с предложенной позволяет улучшить качество полученных вложений.

Ключевые слова: данные, вложение, вектор, функция потерь, линейное пространство, автокодировщик, машинное обучение.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Для цитирования. Ваткин М.Е, Воробей Д.А., Яковлев М.В., Кривова М.Г. Вложенное преобразование с сохранением семантики исходных данных. Доклады БГУИР. 2022; 20(2): 46-52.

EMBEDDING WITH PRESERVATION OF SEMANTICS OF THE ORIGINAL DATA

MAKSIM E. VATKIN, DMITRY A. VOROBEY, MAKSIM V. YAKOVLEV, MARINA G. KRIVOVA

“Sber Bank” (Minsk, Republic of Belarus)

Submitted 27 October 2021

Abstract. In the modern world, the data used to describe objects is often presented as sparse vectors with a large number of features. Working with them can be computationally inefficient, and often leads to overfitting;

therefore, the data dimension reduction algorithms are used, one of which is auto encoders. In this article, we propose a new approach for evaluating the properties of the obtained vectors of lower dimension, as well as a loss function based on this approach. The idea of the suggested loss function is to evaluate the quality of preserving the semantic structure in the embedding space, and to add that metric to loss function to save object relations in the embedding space and thus save more useful information about objects. The results obtained show that using a combination of the mean squared loss function together with the suggested one allows to improve the quality of the embeddings.

Keywords: data, embedding, vector, loss function, linear space, autoencoder, machine learning.

Conflict of interests. The authors declare no conflict of interests.

For citation. Vatkin M.E., Vorobey D.A., Yakovlev M.V., Krivova M.G. Embedding With Preservation of Semantics of the Original Data. Doklady BGUIR. 2022; 20(2): 46-52.

Введение

В банковской сфере обработки данных накапливается большое количество информации о транзакциях клиентов. Такие данные, содержащие информацию о поведении клиента банка, используются банками для построения моделей машинного обучения. Эти данные необходимо трансформировать, чтобы получить их в виде векторов фиксированной длины, где каждая координата действует как счетчик для количества транзакций определенного типа в определенный временной промежуток (например, первая координата показывает количество транзакций на заправочной станции, вторая – в местах общественного питания и т. д.). Такие вектора могут описывать поведение клиента в определенные временные интервалы. Однако ввиду большого количества возможных категорий транзакционной активности эти вектора имеют большое количество координат, многие из которых равны нулю, другими словами, профиль транзакций клиента описывается разряженными векторами.

Использование и хранение таких векторов является вычислительно неэффективным, а также приводит к переобучению при моделировании. Популярным решением данной проблемы является использование автокодировщиков, которые сначала сжимают данные в пространство меньшей размерности, а затем восстанавливают из него изначальные данные. Процесс обучения такой модели может быть описан как сокращение ошибки восстановления путем изменения весов модели, и в результате мы получаем отображение исходных данных в пространство меньшей размерности (вложенное пространство), сохраняющее максимум оригинальной информации об объектах.

Качество полученных представлений измеряется ошибкой восстановления, однако было показано [1], что эта метрика не является надежным индикатором их применимости при решении финальной проблемы. К тому же, она не дает нам представления о том, какую структуру имеет множество точек исследуемого пространства. Это уменьшает уверенность в полученных результатах. Целью данной работы является получить представления, которые содержали бы максимум возможной информации об оригинальных объектах, а также сохраняли их семантические взаимоотношения в эмбеддинговом (вложенном) пространстве. Другими словами, представление объектов в новом пространстве должно быть сформировано так, чтобы при использовании операций сложения и вычитания векторов, мы могли перемещаться от одного представления к другому, и этот переход был бы осмысленным. Это соответствует аналогиям, представленным Миколовым [2], т. е. пары x: y и a: b в исходном пространстве, а x семантически связан с y точно так же, как и a с b, например, «Мужчина»: «Король» и «Женщина»: «Королева», т. е. необходимо, чтобы эта связь была отражена в эмбеддинговом пространстве. Для выполнения этого условия предлагается модифицировать функцию потерь, которая позволяет не только сжать данные в пространство меньшей размерности, но и построить результирующие вектора так, чтобы они отражали семантические взаимоотношения.

Введем обозначения: отображение $f: R^n \to R^m$ $(m<n)$ мы называем эмбеддингом (кодировщиком) и отображение $g: R^m \to R^n$ $(m<n)$ – декодировщиком. Объект представлен

вектором $X \in R^n$, где каждая координата отображает количество совершенных транзакций для продукта/услуги определенной категории.

Цель состоит в построении эмбеддинга, который будет отражать семантические отношения между объектами, что более строго можно сформулировать следующим образом:

$$g(f(x)+f(y)) = z, \quad (1),$$

где z – объект, соответствующий объединению смыслов объектов x и y.

Обращаясь к исходным данным, можно увидеть, что каждый объект сам по себе отражает поведение клиента, более того, существует осмысленная операция сложения исходных данных. Действительно, если мы сложим два исходных вектора, то получим третий вектор, который описал бы третьего клиента, если бы его поведение соответствовало объединению первых двух. Тогда можно переписать (1) как $g(f(x)+f(y)) = x+y$. Если мы предположим, что ошибка восстановления автокодировщика равна нулю, а также f является инъективной функцией, тогда $g(f(x+y))=x+y$ и, следовательно, $f(x)+f(y)=f(x+y)$. Значит, на самом деле (1) говорит о том, что с некоторыми дополнительными условиями f является гомоморфизмом. Похоже, что использование (1) может превратить автокодировщик в метод главных компонент, так как данный метод является линейным отображением, а линейный кодировщик напоминает метод главных компонент [3]. Однако здесь есть допущение, что автокодировщик имеет нулевую ошибку восстановления, что невозможно ввиду разных размеров пространства эмбеддинга и исходного пространства.

Методика проведения эксперимента

Описание данных. В качестве датасета используем датасет, содержащий информацию о 284807 транзакциях за 2 дня от европейских держателей карт в сентябре 2013 года с 492 мошенническими транзакциями (https://www.kaggle.com/mlg-ulb/creditcardfraud). Эти данные использовались ранее исследователями при решении задачи обнаружения мошенничества [4]. Метод состоял в обучении автокодировщика на немошеннических данных и использовании ошибки восстановления, полагая, что мошеннические транзакции будут восстанавливаться автокодировщиком с большей ошибкой восстановления. Поскольку такой датасет ранее часто использовался для исследований, имеется возможность сравнить разрабатываемый подход с аналогичными моделями. Разделим данные на обучающую часть, которая состоит из 80 % транзакций, все из которых нормальные, и тестовую часть, которая содержит оставшиеся 20 % со всеми 492 мошенническими транзакциями. Для того, чтобы использовать предлагаемый в статье подход, необходимо иметь возможность складывать вектора в исходном пространстве, поэтому будем использовать только признаки от V1 до V28.

Метрики обучения. Для измерения процесса обучения используем две метрики:

1) средняя квадратичная ошибка объектов предсказаний автокодировщика для проверки возможностей восстановления автокодировщика;

2) предлагаемая метрика, которую назовем средней ошибкой сохранения семантики. Для ее расчета посчитаем квадратичную ошибку $of\ g(f(x)+f(y))$ и $x+y$, где $x, y \in X$ для каждой пары x и y, и затем усредним. Однако это является вычислительно неэффективным, поэтому выберем для каждого объекта x 25 случайных объектов из X и посчитаем ошибку для них.

Модели. Мы используем три автокодировщика с одинаковой архитектурой: InputLayer(shape=(28,)) – Dense(21, activation='elu') – Dense(14, activation='elu') – Dense(7, activation='elu') – Dense(14, activation='elu') – Dense(21, activation='elu') – Dense(28), но разными функциями потерь.

1. Сумма средней квадратичной ошибки и средней семантической ошибки сохранения (модель 1).

2. Средняя квадратичная ошибка в качестве функции потерь (модель 2).

3. Предложенная функция потерь, которую мы называем средняя семантическая функция потерь. Для ее расчета необходимо выбрать для каждого объекта x из батча n случайных объектов из этого же батча. В результате получим n*batch пар. Затем для каждой пары,

состоящей из объекта x и случайно выбранного объекта y, посчитаем квадратичную ошибку между $g(f(x)+f(y))$ и $x+y$ и возьмем среднее (модель 3).

Размер каждого батча равен 1000, количество эпох – 50, оптимизатор – Adam с параметрами по умолчанию. В качестве предсказаний для задачи обнаружения мошеннических транзакций вычисляем среднюю квадратичную ошибку между входным вектором и его реконструкцией.

Метрики классификации. Так как задача является по своей сути задачей бинарной классификации с дисбалансом классов, используем 3 метрики:

Максимум F1-score. Формула для вычисления: $F1\text{-}score = \frac{2 * precision * recall}{precision + recall}$, где *precision (точность)* – количество объектов, классифицируемых алгоритмом как 1 (мошеннические), и они действительно принадлежат к классу 1, деленное на количество объектов, классифицируемых алгоритмом как класс 1. Формула показывает, можем ли мы доверять алгоритму, когда он относит объект к классу 1; *recall (полнота)* – количество объектов, классифицируемых алгоритмом как 1 (мошеннические), и они действительно принадлежат к классу 1, деленное на количество объектов из класса 1; формула показывает какую долю объектов класса 1 алгоритм может найти. Тем не менее нам нужен баланс этих двух метрик, и для этой цели мы используем *F1-score.* Ввиду того, что мы не знаем какой порог использовать для алгоритма, а эта метрика зависит от порога, мы считаем максимум *F1-score* среди всех возможных.

PRC-AUC – это площадь под кривой точности-полноты, чем ближе эта метрика к 1, тем лучше получилась модель. Для того чтобы построить кривую точности-полноты, сортируются объекты по предсказаниям и выбирается порог так, чтобы он отделял только объект с самым большим предсказанием. Этот объект относят к классу 1, а остальные – к классу 0, после чего находят точность (ось y) и полноту (ось x). Далее порог смещается так, чтобы он отделял теперь два объекта и т. д.

ROC-AUC – это площадь под кривой рабочей характеристики приемника, чем ближе эта метрика к 1, тем лучше получилась модель. Чтобы объяснить, как считается эта метрика, необходимо знать две другие метрики: TPR, которая фактически является полнотой, FPR – количество объектов, классифицируемых алгоритмом как 1 (мошеннические), но они в реальности принадлежат к классу 0, деленное на количество объектов из класса 0. Для того чтобы построить кривую рабочей характеристики приемника, сортируются объекты по предсказаниям и выбирается порог так, чтобы он отделял только объект с самым большим предсказанием. Этот объект относят к классу 1, а остальные – к классу 0, после чего находят TPR (ось y) и FPR (ось x). Далее порог смещается так, чтобы он отделял теперь два объекта и т. д.

Результаты и их обсуждение

Рассмотрим результаты обучения трех моделей (рис. 1, 2, 3), а именно полученные метрики качества во время обучения моделей, метрики классификации, которые показывает качество работы моделей по решению задачи обнаружения мошеннических транзакций, а также их взаимосвязь.

Как видно из рис. 1, модель 2 показывает наименьшую ошибку восстановления (MSE = 0,248 на тренировочных данных, MSE = 0,363 на тестовых данных), ошибка модели 1 больше ошибки модели 2 на 0,04 на тренировочных данных и на 0,046 на тестовых данных (MSE = 0,288 на тренировочных данных, MSE = 0,409 на тестовых данных), а модель 3 показывает наибольшую ошибку восстановления среди трех моделей (MSE = 0,507 на тренировочных данных, MSE = 0,655 на тестовых данных). Тем не менее на рис. 2 модели 1 ($MSPE$ = 0,690 на тренировочных данных, $MSPE$ = 0,910 на тестовых данных) и 3 ($MSPE$ = 0,736 на тренировочных данных, $MSPE$ = 0,936 на тестовых данных) отличаются на 0,046 и 0,026 на тренировочных и тестовых данных соответственно, а модель 2 ($MSPE$ = 1,244 на тренировочных данных, $MSPE$ = 1,473 на тестовых данных) имеет ошибку, большую, чем модель 1, на 0,554 и 0,563 на тренировочных и тестовых данных соответственно.

Из приведенного ранее следует, что комбинация двух функций потерь позволяет получить хорошие результаты по каждой из метрик, тогда как использование только одной приводит к лучшим результатам в соответствующей метрике, но значительно более плохим в другой. Хорошее качество в этих метриках не обязательно должно приводить к хорошим результатам при использовании этих моделей в практической задаче. Модель 2 показала наилучшие качество при восстановлении данных, однако она имеет наихудшие результаты в задаче обнаружения мошеннических транзакций (табл. 1), при этом модель 3 с наихудшим качеством восстановления решает задачу лучше. Лучшая модель также не имеет хорошую характеристику восстановления данных и при этом показывает приемлемое качество по обеим метрикам.

Можно сделать следующие выводы: 1) малая ошибка восстановления в автокодировщике не обязательно ведет к улучшению результатов в практической задаче; 2) с другой стороны, применение комбинированной функции потерь приводит к улучшению финального результата.

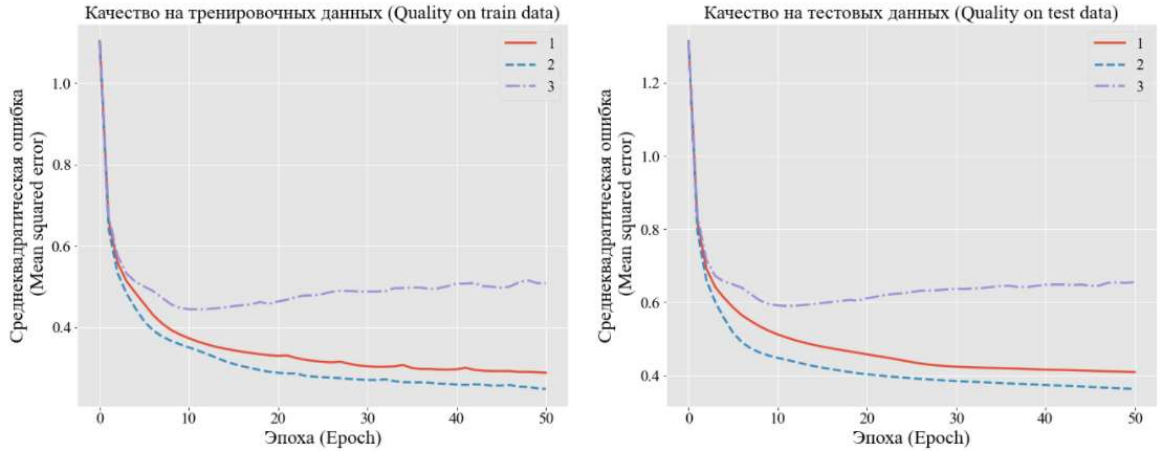

Рис. 1. Ошибка восстановления (1 – модель 1, использует среднеквадратичную ошибку и среднюю семантическую функцию потерь; 2 – модель 2, использует среднеквадратичную ошибку; 3 – модель 3, использует среднюю семантическую функцию потерь)
Fig. 1. Reconstruction loss (1 – model 1, uses MSE and Mean Semantic Preserving Error; 2 – model 2, uses only MSE; 3 – model 3, uses only Mean Semantic Preserving Error)

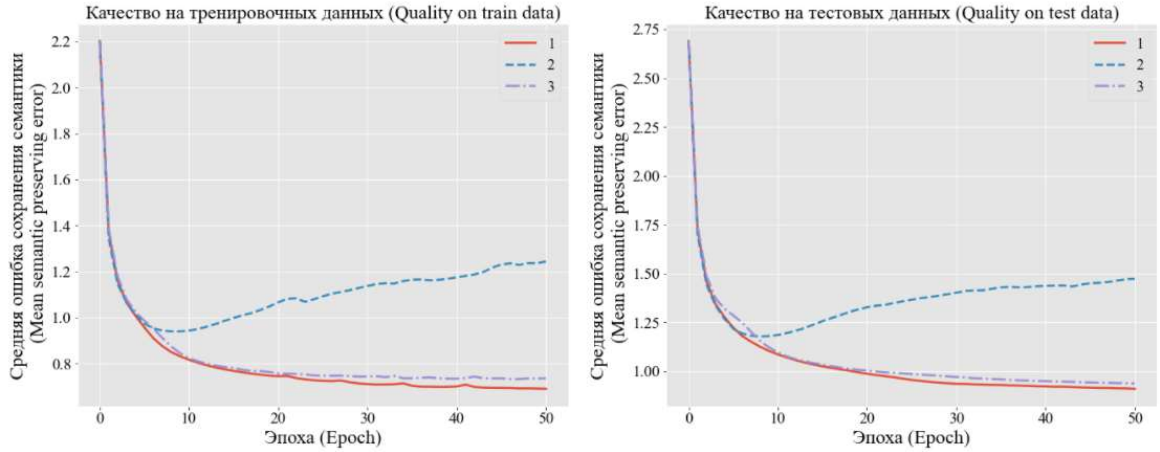

Рис. 2. Средняя ошибка сохранения семантики (1 – модель 1, использует среднеквадратичную ошибку и среднюю семантическую функцию потерь; 2 – модель 2, использует среднеквадратичную ошибку; 3 – модель 3, использует среднюю семантическую функцию потерь)
Fig. 2. Mean Semantic Preserving Error (1 – model 1, uses MSE and Mean Semantic Preserving Error; 2 – model 2, uses only MSE; 3 – model 3, uses only Mean Semantic Preserving Error)

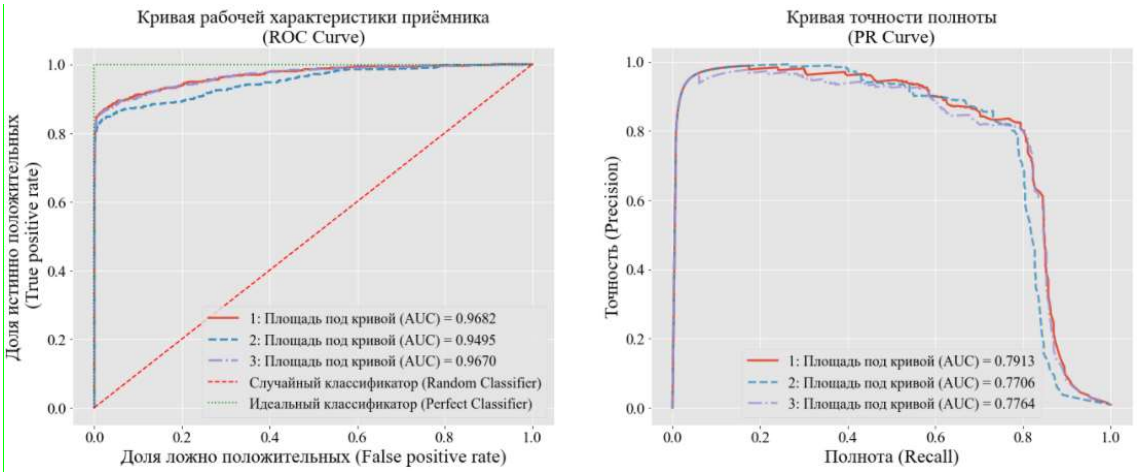

Рис. 3. Кривая рабочей характеристики приемника слева, кривая точности полноты справа (1 – модель 1, использует среднеквадратичную ошибку и среднюю семантическую функцию потерь; 2 – модель 2, использует среднеквадратичную ошибку; 3 – модель 3, использует среднюю семантическую функцию потерь)
Fig. 3. Reveiver operating characteristic curve (ROC) on the left, Precision-Recall Curve on the right (1 – model 1, uses MSE and Mean Semantic Preserving Error; 2 – model 2, uses only MSE; 3 – model 3, uses only Mean Semantic Preserving Error)

Таблица 1. Сравнение метрик классификации
Table 1. Comparison of classification metrics

Алгоритм (algorithm)	Максимальный F1-score (maximum F1-score)	Площадь под PR-кривой (PR-AUC)	Площадь под ROC-кривой (ROC-AUC)
Модель 1 (MSE+MSPE)	0,807851	0,788064	0,968177
Модель 2 (MSE)	0,801644	0,772925	0,967016
Модель 3 (MSPE)	0,794148	0,770675	0,949515

Проиллюстрируем качество полученного решения относительно результатов, полученных другими исследователями. Сравнение проведем, основываясь на метрике PR AUC, так как данная метрика лучше, чем метрика ROC AUC для измерения качества в задачах с дисбалансом классов [5]. Для сравнения возьмем модели из статьи [6]. Из табл. 2 можно увидеть, что модель 1 (MSE+MSPE) заняла второе место, т. е. предложенный подход сравним с другими существующими алгоритмами для решения задач такого рода.

Таблица 2. Сравнение площади под графиком кривой точности-полноты с другими алгоритмами
Table 2. PR-AUC comparison with other algorithms

Алгоритм / (algorithm)	Площадь под PR-кривой (PR-AUC)
Бэггинг	0,825
Модель 1(MSE+MSPE)	0,788
C4.5	0,745
Наивный Байес	0,080

Заключение

В статье представлена модификация типичной функции потерь для автокодировщика, которая может использоваться для сохранения семантических отношений между объектами в эмбеддинговом пространстве для определенного типа табличных данных. Однако могут быть дополнительные свойства результирующего пространства и его отличия от обычного подхода, которые еще предстоит изучить.

Список литературы / References

1. Gupta P., Banchs R.E., and Rosso P. Squeezing bottlenecks: exploring the limits of autoencoder semantic representation capabilities. *Neurocomputing*. 2016;175:1001–1008.

2. Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J. Distributed representations of words and phrases and their compositionality. *NIPS. 2013:*3111–3119.
3. Bourlard H., Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition. *Biol. Cybern.* 1988;59(September (4)):291-294. DOI: 10.1007/bf00332918.
4. Al-Shabi M.A. Credit Card Fraud Detection Using Autoencoder Model in Unbalanced Datasets. *JAMCS.* 2019;33(5):1-16.
5. Saito T., Rehmsmeier M. The Precision-Recall Plot is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One*. 2015;10(3).
6. Husejinović A. Credit card fraud detection using naive Bayesian and C4.5 decision tree classifiers. *Periodicals of Engineering and Natural Sciences.* 2020;8(1):1-5.

Вклад авторов

Воробей Д.А. формализовал идею статьи и привел ее к конечному виду, провел эксперименты по реализации идеи в виде функции потерь.

Ваткин М.Е. подготовил данные для проведения эксперимента, предложил способ сравнения результатов предложенного подхода с классическим подходом.

Яковлев М.В. предложил перенести идеи из обработки естественного языка на табличные данные, выполнил интерпретацию результатов экспериментов.

Кривова М.Г. провела критическую оценку используемого подхода и результатов экспериментов, оказала помощь в подготовке текста статьи.

Authors' contribution

Vorobey D.A. formalized the idea of the article and brought it to its final form, as well as conducted experiments to implement the idea in the form of a loss function.

Vatkin M.E. prepared the data for the experiment, as well as suggested a way to compare the results of the approach used with the classical approach.

Yakovlev M.V. proposed to transfer ideas from natural language processing to tabular data, and was engaged in the interpretation of experimental results.

Krivova M.G. carried out a critical assessment of the approach used and the results of experiments, as well as assisted in the preparation of the content of the article.

Сведения об авторах

Ваткин М.Е, к.т.н., главный специалист по данным ОАО «Сбер Банк».

Воробей Д.А, специалист по данным ОАО «Сбер Банк».

Яковлев М.В, специалист по данным ОАО «Сбер Банк».

Кривова М.Г., специалист по данным ОАО «Сбер Банк».

Information about the authors

Vatkin M.E, Cand. of Sci., Chief Data Scientist of "Sber Bank".

Vorobey D.A, Data Scientist of "Sber Bank".

Yakovlev M.V, Data Scientist of "Sber Bank ".

Krivova M.G., Data Scientist of "Sber Bank ".

Адрес для корреспонденции

220005, Республика Беларусь,
г. Минск, Бульвар Мулявина 6,
ОАО «Сбер Банк»;
тел. +375-29-278-13-78;
e-mail: mevatkin@bps-sberbank.by;
Ваткин Максим Евгеньевич

Address for correspondence

220005, Republic of Belarus,
Minsk, Mulyavina blv., 6,
«*Sber Bank*»;
tel. +375-29-278-13-78;
e-mail: mevatkin@bps-sberbank.by;
Vatkin Maksim Evgenyevich