УДК 004.896

ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ И PROCESS MINING К ЗАДАЧАМ КЛАСТЕРИЗАЦИИ ПРОЦЕССОВ В ФИНАНСОВОЙ СФЕРЕ

КРЕВСКИЙ МИХАИЛ ИГОРЕВИЧ

Кафедра банковских информационных технологий, Физтех-школа Прикладной Математики и Информатики, ФГАОУ ВО «Московский физико-технический институт (НИУ)», Институтский пер., д.9, 141701, г.Долгопрудный, Московская область, Российская Федерация, e-mail: kremi999@mail.ru, телефон: +7 (965) 209-11-14

СМЕТАНЕВ ДАНИИЛ МИХАЙЛОВИЧ

Центр Компетенций Process Mining, Департамент финансов, ПАО «Сбербанк России», ул. Вавилова, д.19, 117997, Москва, Российская Федерация, e-mail: danil-smetanev@yandex.ru, телефон: +7 (915) 470-42-42

Аннотация

Важным инструментом повышения эффективности бизнес-процессов в компаниях может стать process mining, являющийся промежуточным звеном между data mining и process management. Process mining – это совокупность методов и подходов к извлечению, анализу и оптимизации процессов на основании изучения данных из журналов событий (eventlogs) информационных систем. Она приходит на смену классическим методам описания и ручного моделирования процессов и позволяет из необработанных цифровых данных получать информацию о связанных с ними процессах. Применение машинного обучения в связке с process mining является крайне перспективным направлением. В статье показано использование этой связки для решения задачи кластеризации журнала событий. Рассмотрено три различных способа векторизации процессов в журнале: label encoding, one-hot-encoding и act2vec. После была проведена кластеризация получившихся векторов с помощью одного и того же алгоритма k-means. В результате был проведен сравнительный анализ трех способов. Label encoding проявил себя неудовлетворительно, показав крайне низкое качество на этапе кластеризации. One-hot-encoding продемонстрировал значительно лучший результат и справился с кластеризацией, но число столбцов в свойственной этой векторизации разреженной матрице будет резко расти при увеличении количества вариантов у категориальных признаков. Лучшим решением оказался act2vec – аналог word2vec из обработки естественного языка, примененный к процессам из журнала событий. Act2vec, как и его близкий родственник, является сравнительно экономичным отображением в признаковое пространство фиксированной размерности. По качеству он не уступил one-hot-encoding, но обошел его по производительности.

Ключевые слова: машинное обучение; process mining; data science; векторизация процессов; кластеризация; обработка естественного языка

Введение

Перевод на цифровую основу является одним из наиболее мощных инструментов повышения конкурентоспособности современного бизнеса. Важным инструментом повышения эффективности бизнес-процессов в компаниях может стать process mining [1], являющийся промежуточным звеном между data mining и process management. Process mining – это совокупность методов и подходов к извлечению, анализу и оптимизации процессов на основании изучения данных из журналов событий (eventlogs) информационных систем. Она приходит на смену классическим методам описания и ручного моделирования процессов. Восстанавливая схемы бизнес-процессов в том виде, как они протекают на самом деле,

process mining позволяет получить полную информацию, необходимую для устранения «бутылочных горлышек» модели – узких мест, оптимизация которых ведет к значительному повышению производительности.

Функционирование process mining основано на использовании данных журналов событий (eventlogs). Информационные системы автоматически регистрируют все процессы (activities), выполненные в ходе одной цепочки событий. В дополнение к виду деятельности журналы событий содержат временные метки (times tamps), уникальный идентификационный номер (case ID) и прочие атрибуты события: пользователи процесса, текст переписки, решение по сделке и т.д. С помощью этой информации можно визуализировать и проанализировать модель реального, а не предполагаемого процесса. Именно это и делает process mining. Однако он не ограничивается восстановлением моделей бизнес-процессов из журналов событий. Сопоставляя данные журналов и схемы процессов, process mining позволяет выявить отклонения и узкие места в цепочке процессов, изучить длительность выполнения каждой ее стадии, обнаружить лишние или недостающие этапы, а также определить потенциальные способы оптимизации и, следовательно, повышения эффективности.

В настоящее время одной из ключевых задач, стоящих перед компаниями, является извлечение скрытых зависимостей из данных, имеющихся в корпоративных системах. Владея цифровой информацией обо всех действиях, выполненных в рамках процесса, с помощью технологии process mining компания способна воспроизвести фактически исполняемый бизнес-процесс. А построение его визуальной схемы позволяет не только получить полное представление о цепочке событий, но и исследовать актуальное состояние процесса на любом уровне детализации. Анализируя модель реконструированного процесса вместе с данными о длительности и особенностях его исполнения, можно выявить задержки по времени реализации отдельных действий, взаимосвязи между пользователями, зацикленности в процессе, неэффективных исполнителей, а также скрытые недостатки и проблемы в процессах, из-за которых может существенно снижаться производительность целой организации. Помимо этого process mining позволяет увидеть процесс в динамике, а это означает возможность отследить изменения, происходящие в результате введения тех или иных мер по усовершенствованию. Итак, process mining играет ключевую роль в обеспечении компаний ценной и объективной информацией, необходимой для повышения эффективности решений по оптимизации бизнес-процессов.

Основная часть

В статье рассмотрены и подвергнуты сравнительному анализу три вида векторизации процессов: lable encoding, one-hot-encoding и act2vec. Входные данные для векторизации представлены в виде журнала событий. Полученные виды векторных представлений кластеризованы с помощью алгоритма кластеризации k-means. Далее выполнен сравнительный анализ кластеров, на основе которого выбран лучший из рассмотренных методов векторизации процессов.

Из указанных видов векторизации наибольший интерес для автора представляет act2vec [2], поэтому приводится его более подробное описание. Входные данные представлены в виде журнала событий. При работе с процессами возможно использовать подход, аналогичный подходу word2vec [3] в обработке естественного языка (NLP). Процессы можно рассматривать как слова в контексте предложения, но здесь контекстом является журнал событий одного пользователя. Журнал событий пользователя в представленной задаче будет являться следом (trace) в терминологии process mining и аналогом предложения в NLP. Несмотря на то, что следы в журнале событий являются последовательными по своей природе, определять контекст процесса можно и на основе неупорядоченного набора предшествующих и последующих действий. Таким образом, получается представление, которое может быть использовано для понимания контекста двух разных видов процессов. В зависимости от их контекста их представление будет либо одинаковым, либо очень различным.

Кластеризация – это задача разбиения множества объектов на группы. Внутри каждой группы, называемой кластером, должны оказаться похожие объекты, а объекты разных кластеров должны быть как можно более отличны. Для кластеризации работает способ обучения без учителя, то есть для обучения модели не требуется заранее размеченный датасет. В задаче был применен классический алгоритм кластеризации k-means. В качестве метрики качества можно использовать сумму квадратичных ошибок (SSE) [4].

В качестве объекта кластеризации – датасета – будет использоваться сгенерированный журнал событий. В дальнейшем в него будут вноситься изменения, и для ясного понимания их логики необходимо описать исходный датасет (Рис. 1).

	Prossess id	Activity	User	Timepoint
0	1	a	u2	15.04.2016 8:41
1	1	b	u3	18.04.2016 12:55
2	2	a	u2	18.04.2016 20:40
3	1	d	u5	19.04.2016 7:22
4	2	b	u4	21.04.2016 22:42
...	...	...	...	...
45374	7549	e	u7	20.07.2020 18:29
45375	7547	f	u7	21.07.2020 23:42
45376	7549	f	u7	22.07.2020 3:43
45377	7547	h	u1	22.07.2020 22:17
45378	7549	h	u2	26.07.2020 11:17

45379 rows × 4 columns

Рис. 1. Неизмененный датасет.

Процесс имеет четыре признака: идентификатор процесса, обозначение вида процесса, обозначение пользователя и временная метка. Идентификатор процесса отмечает порядковый номер процесса у определенного пользователя и смысла для анализа не несет, его можно сразу удалить. Обозначение вида процесса и обозначение пользователя являются категориальными признаками. Временная метка хранится в виде строки, чтобы с ней взаимодействовать, лучше воспользоваться функцией из библиотеки time - strptime и сменить формат на struct_time.

1. Label encoding

Для чистоты эксперимента имеет смысл запустить кластеризацию на минимально измененном датасете. Чтобы модель смогла обрабатывать данные категориальных признаков, необходимо хотя бы преобразовать каждое значение категориального столбца в соответствующее ему число. Из признака времени пока можно оставить только день недели, так как модель не сможет работать со значениями типа struct_time. Struct_time – структура из библиотеки time, характеризующая время, откуда можно извлечь всю необходимую информацию. Она содержит поля tm_year – номер года, tm_mon – номер месяца, tm_mday – число месяца, tm_hour – часов от начала суток, tm_min – минут от начала часа, tm_sec – секунд от начала минуты, tm_wday – день недели и tm_yday – номер дня в году. Чтобы улучшить результат, также можно попробовать добавить столбец с рассчитанным временем между соседними действиями пользователя. При использовании предлагаемой модели целесообразно перемешивать строки в случайном порядке для избежания возможных временных зависимостей. Датасет для label encoding представлен на рис. 2.

	Activity	User	Week_day	Dif_time
16949	1	3	2	659460.0
42969	3	5	1	1469340.0
24474	4	7	2	1120020.0
35043	1	3	0	1147920.0
41437	3	5	0	383160.0
...	...	...	...	...
20970	6	5	1	268860.0
2931	5	8	3	105960.0
21489	1	3	5	476880.0
43626	6	6	1	233160.0
8239	0	2	4	788880.0

45379 rows × 4 columns

Рис. 2. Датасет для label encoding.

Рассмотрев экспериментально зависимости метрики качества кластеризации от количества кластеров, выберем лучший результат. Оптимальным числом кластеров оказалось 4. Как видно из графиков на рис. 3-6, при label encoding кластеризация выдает неудовлетворительный результат. Точки не удалось правильно разделить.

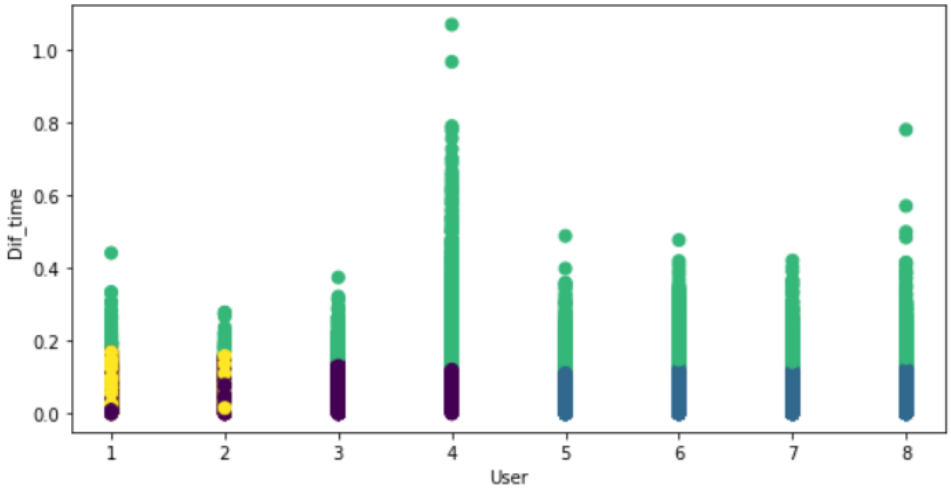

Рис. 3. Датасет, кластеризированный с помощью labelencoding, в координатах времени между процессами и номера пользователя.

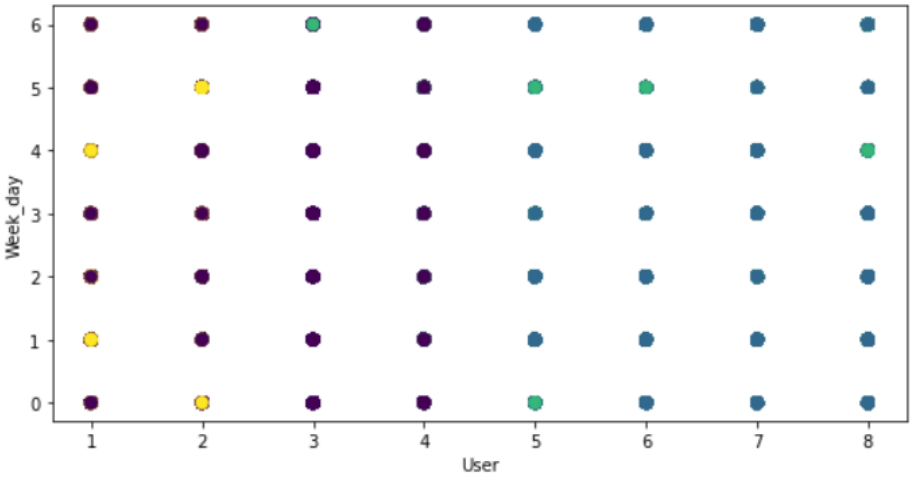

Рис. 4. Датасет, кластеризированный с помощью labe encoding, в координатах дня недели проведения процесса и номера пользователя.

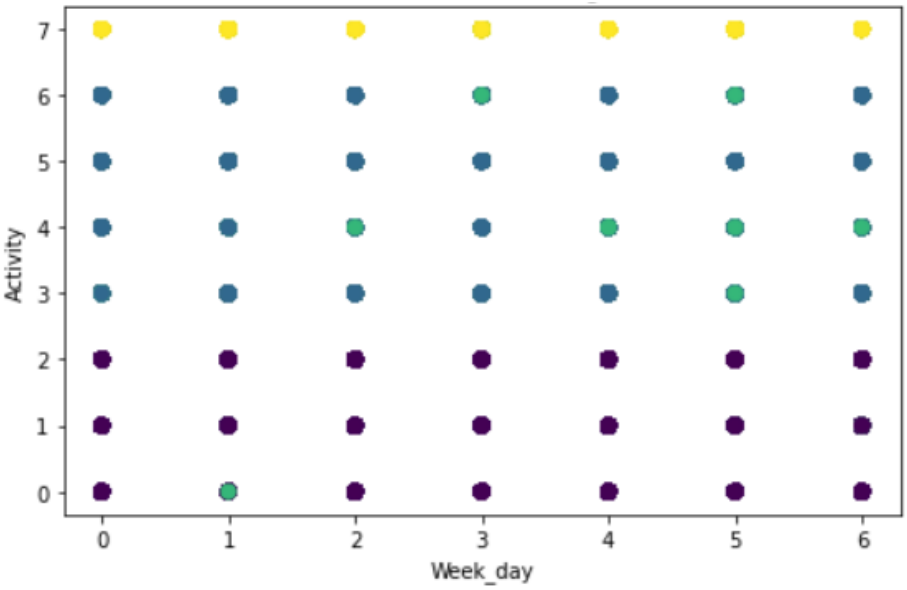

Рис. 5. Датасет, кластеризированный с помощью label encoding, в координатах вида процесса и номера пользователя.

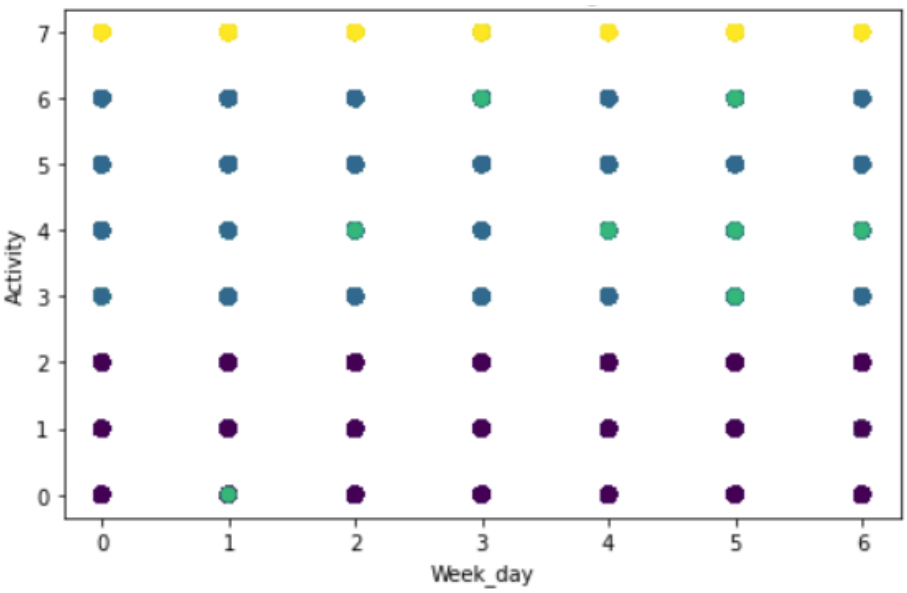

Рис. 6. Датасет, кластеризированный с помощью label encoding, в координатах вида процесса и дня недели проведения процесса.

Это является следствием того, что, присваивая категориальным признакам числовые значения, мы порождаем новые, ненужные зависимости внутри них, поэтому значительно снижается качество результата. Например, получается, что пользователь u4 в два раза больше пользователя u2.

2. One-hot-encoding

Простейший способ решения проблемы категориальных признаков – это преобразовать их, использовав one-hot-encoding. Если один категориальный признак User может принимать 8 значений, то после преобразования получится 8 отдельных признаков, где в строке есть только одна единица, а все остальные значения – нули. Так из одного столбца выходит разреженная матрица с 0 и 1. Подобный способ позволяет добиться значительного улучшения качества. В результате одновременно имеются признаки, которые легко будут обрабатываться моделью машинного обучения, и не создаются новые, ложные зависимости и информация. Например, пользователи оказываются просто либо равны друг другу, либо не равны. Для реализации преобразований воспользуемся функцией pandas.get_dummies. Теперь снова стоит перемешать датасет (Рис. 7) и запустить кластеризацию.

	Week_day	Dif_time	User_1	User_2	User_3	User_4	User_5	User_6	User_7	User_8	Activity_0	Activity_1	Activity_2	Activity_3	Activity_4	Activity_5	Activity_6	Activity_7
578	5	26280.0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
30491	4	499860.0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
17777	1	552600.0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
27148	2	776220.0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7536	3	463980.0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
29415	5	435480.0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
9539	5	390720.0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
12532	4	443820.0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0
190	3	250800.0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
44392	3	430680.0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0

45379 rows × 18 columns

Рис. 7. Разреженный датасет для one-hot-encoding.

Выполнив эксперимент, получаем, что оптимальное число кластеров при one-hot-encoding – тоже 4. И по метрике, и визуально кластеризация с one-hot-encoding значительно превосходит таковую с label encoding, что отображено на рис. 8-11.

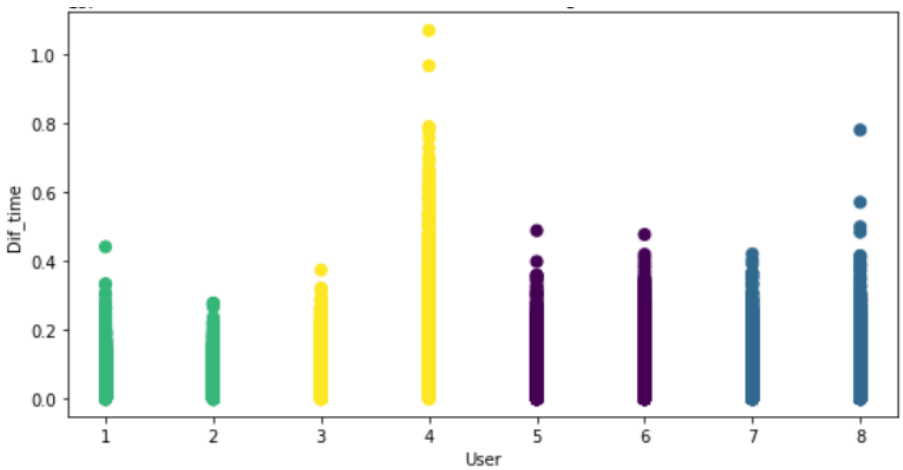

Рис. 8. Датасет, кластеризированный с помощью one-hot-encoding, в координатах времени между процессами и номера пользователя.

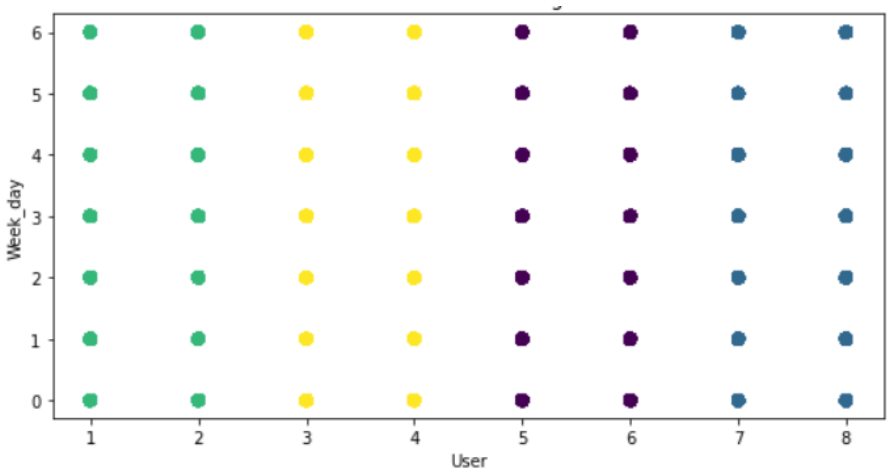

Рис. 9. Датасет, кластеризированный с помощью one-hot-encoding, в координатах дня недели проведения процесса и номера пользователя.

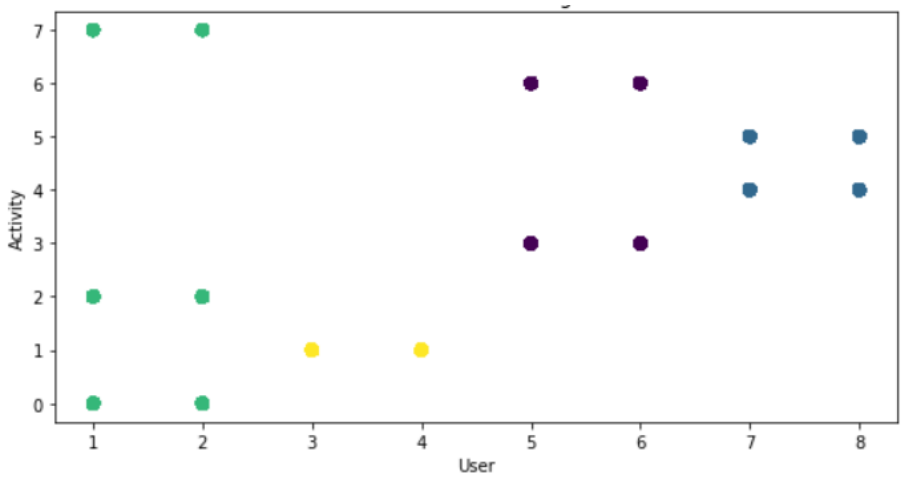

Рис. 10. Датасет, кластеризированный с помощью one-hot-encoding, в координатах вида процесса и номера пользователя.

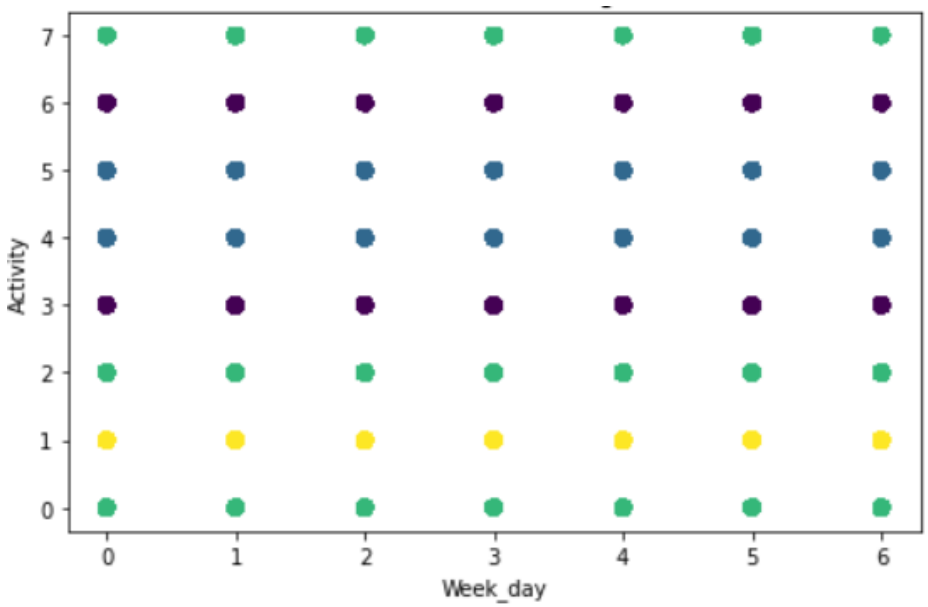

Рис. 11. Датасет, кластеризированный с помощью one-hot-encoding, в координатах вида процесса и дня недели проведения процесса.

3. Act2vec

В третьем, самом интересном из рассмотренных подходов к векторизации – контекстном act2vec, используется прямая аналогия с векторизацией в word2vec. Имеет смысл напомнить о преимуществах word2vec перед one-hot-encoding в обработке естественного языка. В работе с бизнес-процессами преимущества act2vec будут носить точно такой же характер.

Как было написано выше, при one-hot-encoding для каждого категориального признака генерируется разреженная матрица, где в каждой строке есть лишь одна единица. Причем ширина этой матрицы будет равна количеству значений, которые может принимать этот категориальный признак. В уже имеющемся датасете признак user продуцирует 8 столбцов, а признак activity – 7. Очевидно, когда в журнале событий будет больше полей, а количество пользователей и видов действий превысит несколько сотен, датасет станет слишком большим и неэффективно тяжеловесным для использования в машинном обучении.

В свое время эта проблема была одной из главных причин, подтолкнувших к развитию word2vec. В act2vec сжатые векторные представления строятся в пространствах фиксированной размерности. Кроме того, такие представления для похожих процессов близки как векторы.

В конкретной задаче нужно с помощью нейросети заполнить матрицу эмбеддингов. Из этой нормированной матрицы весов извлекаются векторы, являющиеся контекстной репрезентацией процессов. Получившиеся векторы остается только кластеризовать.

Графики, отражающие кластеризированный датасет при act2vec, абсолютно идентичны таковым при one-hot-encoding. Из этого и по метрике заметно, что act2vec ничуть не уступает по качеству one-hot-encoding, но при увеличении признакового пространства даст значительный выигрыш в производительности.

Заключение

В статье рассмотрены и подвергнуты сравнительному анализу три вида векторизации процессов: lable encoding, one-hot-encoding и act2vec. На основе кластеризации векторных представлений входных данных с помощью алгоритма кластеризации k-means, можно утверждать, что в задаче кластеризации методы act2vec и one-hot-encoding показывают значительно лучшее качество, чем lable encoding. С экспериментальным датасетом невозможно показать, что act2vec значительно экономичнее one-hot-encoding. Но это очевидно из теоретических соображений и будет проявляться при увеличении размерности признакового пространства. Таким образом, в результате проведенного эксперимента получен вывод, что наилучшим из рассмотренных методов векторизации процессов можно считать act2vec.

Список литературы

[1] Pieter De Koninck. Act2vec, trace2vec, log2vec, model2vec: Representation Learning for Business Processes

[2] Aurelien Geron. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow

[3] Wil van der Aalst, Coursera. Process Mining: Data science in Action

[4] Tomas Mikolov, Distributed Representations of Words and Phrasesand their CompositionalityTomas

APPLICATION OF MACHINE LEARNING AND PROCESS MINING TO THE PROCESS CLUSTERING IN FINANCIAL SPHERE

KREVSKIY MIKHAIL IGOREVICH

Department BIT, Phystech-School of Applied Mathematics and Computer Science, Moscow Institute of physics and technology (NIU), 9 Institutsky per., 141701, Dolgoprudny, Moscow region, Russian Federation, e-mail: kremi999@mail.ru, +7 (965) 209-11-14

SMETANEV DANIIL MIKHAILOVICH

Process Mining Competention Center, Financial Department, PJSC Sberbank, 19 Vavilova str., 117997, Moscow, Russian Federation, e-mail: danil-smetanev@yandex.ru, +7 (915) 470-42-42

Annotation

Data science is a branch of computer science related to the analysis, processing and presentation of data. Two opposite directions can be distinguished in this science: data-oriented analysis and process analysis. The first direction includes data mining and machine learning, the second - simulation, verification and optimization of processes. Between them there is process mining, which allows you to obtain information about the processes from raw digital data. The use of machine learning in conjunction with process mining is an extremely promising area. The article discusses the use of this bundle to solve the problem of clustering the event log. Three different ways of vectorizing processes in the log are considered: label encoding, one-hot-encoding and act2vec. After that, the resulting vectors were clustered using k-means algorithm. As a result, a comparative analysis of three methods was carried out. Label encoding was unsatisfactory, showing extremely low quality at the clustering stage. One-hot-encoding showed a significantly better result and coped with clustering, but the number of columns in related to this vectorization sparse matrix will increase dramatically with an increase in number of categorical attributes. Act2vec turned out to be the best solution - an analogue of word2vec from natural language processing applied to processes from the event log. Act2vec, like its close relative, is a relatively cheap and effective mapping into a feature space of a fixed dimension. It was not inferior in quality to one-hot-encoding, but bested it in performance.

Key words: machine learning; process mining; data science; process vectorization; clustering; natural language processing