Дискретизация непрерывной величины, характеризующей интенсивность, в модели социально-значимого поведения

А. В. Торопова
Санкт-Петербургский Федеральный исследовательский центр РАН
alexandra.toropova@gmail.com

Т. В. Тулупьева[1,2,3]
[1]*Санкт-Петербургский Федеральный исследовательский центр РАН;*
[2]*Северо-Западный институт управления РАНХиГС;*
[3]*Санкт-Петербургский государственный университет*
tvt@dscs.pro

***Аннотация*. Интенсивность – это одна из основных характеристик поведения, определить которую можно, как среднее число эпизодов поведения, случившихся за определенный период времени. Знание интенсивности поведения может быть использовано во многих прикладных областях для прогнозирования поведения и оценки других, связанных с ним свойств. Ранее была представлена модель на основе байесовской сети доверия, позволяющая оценить интенсивность поведения, используя данные о последних эпизодах поведения, а также минимальном и максимальном интервалах между эпизодами. В связи с тем, что байесовские сети доверия предполагают работу с дискретными величинами, в модели используется дискретизация непрерывных величин. В данной работе впервые рассматривается вопрос того, как различные методы дискретизации непрерывной величины, характеризующей интенсивность поведения, влияют на эффективность предсказаний этой модели.**

Ключевые слова: интенсивность поведения; байесовская сеть доверия; дискретизация непрерывных величин; эпизоды поведения

I. Введение

Моделирование поведения человека уже долгое время остаётся одной из наиболее актуальных задач. При этом зачастую приходится иметь дело с различного рода неопределённостями, связанными с недетерминированностью поведения и сложностью его предсказания. Само же поведение может быть охарактеризовано множеством параметров, одним из важнейших параметров является интенсивность.

Интенсивность – это одна из основных характеристик поведения, определить которую можно, как среднее число эпизодов поведения, случившихся за определенный период времени [1]. Имея данные об интенсивности некоторого поведения, можно делать прогнозы об интенсивности этого поведения в будущем, что в свою очередь может быть использовано в различных областях, включая здравоохранение, образование, социологию. Например, по интенсивности определенных действий в социальных сетях (публикация постов, лайки и т.д.) можно определить эмоциональное состояние человека или возможность его участия в социоинженерных атаках [2, 3], другой пример: по частоте определенных действий медперсонала можно предсказать концентрацию вируса на их руках [4].

Применение байесовских сетей доверия подразумевает использование дискретных величин. Целью данной работы является повышение эффективности оценки интенсивности поведения модели социально-значимого поведения за счет наиболее подходящего выбора дискретизации непрерывных величин, входящих в модель. Дискретизация определяется как процесс разбиения непрерывной переменной на различные категории в зависимости от того, в какой интервал она попадает [5]. Дискретизация зачастую является одним из важнейших этапов предварительной обработки данных [6].

Новизна данного исследования заключается в том, что впервые рассмотрено влияние различных вариантов дискретизации непрерывных величин на работу модели оценки интенсивности поведения. Так как дискретизация может значительно повлиять на эффективность работы модели [7], рассмотрение данного вопроса подготовит базу для дальнейшего использования и изучения модели.

II. Описание модели

На рис. 1 представлена структура модели оценки интенсивности поведения [8]. Модель представляет собой байесовскую сеть доверия. Тензоры условной вероятности, определяющие переходы между узлами сети, высчитываются с помощью машинного обучения на основе синтетических данных.

Вершина λ характеризует интенсивность поведения, *t_next* – интервал между последним эпизодом за исследуемый период и первым эпизодом после окончания исследуемого периода, *t_12* – интервал между последним и предпоследним эпизодами поведения, *t_23* – интервал между предпоследним и третьим с конца эпизодами поведения за исследуемый период, *t_min* и *t_max* – минимальный и максимальный интервалы между эпизодами за исследуемый период, *n* – количество эпизодов за исследуемый период.

Данные для обучения и тестирования модели были синтезированы автоматически по алгоритму, описанному в [9], специально для этого была написана программа на

Работа выполнена в рамках проекта по государственному заданию СПб ФИЦ АН № 0073-2019-0003, при финансовой поддержке РФФИ, проекты №19-37-90120, № 20-07-00839

языке R [10] с использованием пакета bnlearn [11] для работы с байесовскими сетями доверия. Анализ был также выполнен с помощью языка R. Итоговый обучающий набор содержит 6000 записей. Отметим, что при этом задано исходное значение интенсивности. Благодаря этому возможно его сравнение с предсказанием модели.

Последующее тестирование модели проводится на синтезированных аналогичным образом тестовых данных, которые содержат 200 значений интенсивности, для каждого значения — 10 последовательностей, в совокупности 2000 записей.

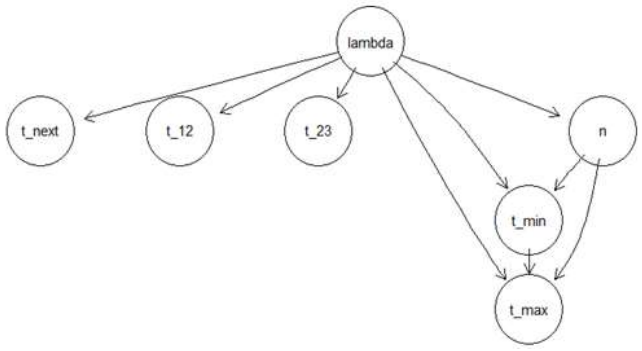

Рис. 1. Модель интенсивности поведения со скрытыми переменными

III. Дискретизация непрерывных величин

Алгоритмы дискретизации можно разделить на два вида: контролируемые (supervised) и неконтролируемые (unsupervised) [12–14]. Для применения контролируемого алгоритма дискретизации экземпляры выборки должны принадлежать к какому-то классу. А если есть только множество значений, которые нужно разделить на интервалы, используется неконтролируемый алгоритм. Заметим, что неконтролируемые алгоритмы можно применять в обоих случаях. В последнее время большинство исследований посвящено на создание и совершенствование контролируемых алгоритмов [15], однако неконтролируемые алгоритмы не могут быть использованы при решении задачи дискретизации интенсивности поведения, так как для нее нет какой-то определенной классификации.

Следовательно, в связи с отсутствием принадлежности к какому-то определённому классу для дискретизации случайной величины, характеризующей интенсивность процесса, можно применить только неконтролируемые алгоритмы. Для дискретизации временных интервалов могут быть использованы обе группы алгоритмов, однако ее логично сделать такой, чтобы она отражала уже устоявшиеся обозначения временных промежутков, такие как час, день, полдня, неделя, месяц, полгода, год и т. д. Так как данные об интервалах между эпизодами поведения могут быть получены основном из опросов респондентов, есть большая вероятность, что именно такие обозначения и будут использованы в их ответах. За одну единицу времени был взят один день, и для случайных величин *t_i,j+1*, *t_next*, *t_min* и *t_max* была взята дискретизация вида: $t^{(1)}=[0;0.1)$, $t^{(2)}=[0.1;0.5)$, $t^{(3)}=[0.5;1)$, $t^{(4)}=[1;7)$, $t^{(5)}=[7;30)$, $t^{(6)}=[30;90)$, $t^{(7)}=[90;180)$, $t^{(8)}=[180;\infty)$, то есть к первому интервалу относятся интервалы примерно до двух с половиной часов, ко второму — от двух с половиной часов до половины дня, к третьему — от половины дня до одного дня, к четвертому — от одного дня до недели, к пятому — от недели до месяца, к шестому от месяца до примерно трех месяцев, к седьмому – от трех месяцев до полугода, к восьмому — большие, чем полгода. Для дискретизации случайной величины, характеризующей интенсивность поведения, были применены следующие алгоритмы: разбиение на равные по величине отрезки (EW), разбиение на интервалы равные по частоте (EF) и модифицированное разбиение на интервалы равные по частоте (EF_Unique) [13]. Рассмотрим эти алгоритмы подробнее.

EW (Equal width) – это один из самых популярных и простых методов дискретизации разделяет диапазон значений ($R=max-min$) данных непрерывной величины, отсортированных в восходящем порядке, на k равных интервалов с $k–1$ точками разрыва ($c1, c2, \ldots c(k–1)$), вычисляемых по формуле $ci=min+i\cdot h$, $i=1,\ldots,k–1$. Длина интервалов (h) определяется как частное от деления диапазона значений на количество интервалов k: $h=R/k$. Проблема EW может заключаться в том, что в некоторых интервалах может вообще не оказаться никаких значений, а в некоторых — наоборот будет содержаться намного больше, чем в других интервалах.

EF (Equal frequency) является легкореализуемым и доступным алгоритмом. Суть его работы заключается в разделении диапазона значений на k интервалов, каждый из которых содержит примерно n/k значений, где n – это количество всех значений. Алгоритм заключается в следующем: все значения сортируются в порядке возрастания и делятся на k групп, точки разрыва определяются как средние арифметическое максимального значения текущей группы и минимального значения следующей группы, после этого все непрерывные значения переводятся в значения интервалов, в которых они содержатся. К преимуществам данного метода можно отнести то, что похожие значения будут собраны в одном интервале, кроме того, эффект резко отличающихся значений, который можно увидеть при применении EW, будет снижен. Недостаток метода состоит в том, что два и даже больше близлежащих интервалов могут содержать одинаковые значения [14].

EF_Unique – это модифицированный алгоритм EF. EF_Unique состоит из двух основных шагов: на первом шаге все значения сортируются в восходящем или убывающем порядке, затем удаляются дублирующие значения, число интервалов k устанавливается как ближайшее целое число к квадратному корню из числа уникальных значений, которые будут дискретизированы. Таким образом этот метод может разбивать различные атрибуты наборов данных на разные количества интервалов. На втором шаге оставшиеся значения разделяются на k групп, соответственно методу EF. Затем высчитываются средние арифметические этих групп, чтобы определить границы интервалов, после этого точки разрыва вычисляются как средние арифметические находящихся друг за другом ячеек. И наконец непрерывные значения атрибута переводятся в дискретные с помощью определения интервала, в котором они находятся.

Выбор оптимального числа количества интервалов k – это еще одна важная задача, связанная с

дискретизацией непрерывных величин. При небольшом *k* может быть утрачена некоторая часть информации, а при больших значениях *k* может оказаться довольно сложно правильно интерпретировать полученные результаты [14]. Для текущего исследования рассмотрим 4 значения *k*: 7, 9, 10, 12.

В EF_Unique число интервалов высчитывается в самом алгоритме, однако рассмотрим дискретизацию для всех четырех вариантов.

Так как значения интенсивности могут изменяться от 0 до бесконечности, дискретизацию для всех интервалов будем начинать с 0 и заканчивать бесконечностью.

Результаты дискретизации получились следующими (при округлении до тысячных).

При использовании EW точки разрыва при *k=7 – {0.232, 0.463, 0.695, 0.926, 1.157, 1.389}*, при *k=9 – {0.181, 0.361, 0.541, 0.721, 0.9, 1.08, 1.26, 1.44}*, при *k=10 – {0.163, 0.325, 0.487, 0.649, 0.81, 0.972, 1.134, 1.296, 1.458}*, при *k=12 – {0.136, 0.271, 0.406, 0.541, 0.676, 0.81, 0.945, 1.08, 1.215, 1.35, 1.485}*.

При использовании EF точки разрыва при *k=7 – {0.055, 0.116, 0.21, 0.317, 0.471, 0.678}*, при *k=9 – {0.046, 0.097, 0.149, 0.217, 0.306, 0.393, 0.552, 0.769}*, при *k=10 – {0.042, 0.083, 0.125, 0.198, 0.244, 0.332, 0.442, 0.595, 0.791}*, при *k=12 – {0.035, 0.07, 0.106, 0.149, 0.207, 0.244, 0.32, 0.393, 0.506, 0.644, 0.835}*.

При использовании EF_Unique точки разрыва при *k=7 – {0.061, 0.129, 0.212, 0.318, 0.475, 0.793}*, при *k=9 – {0.048, 0.096, 0.156, 0.228, 0.309, 0.424, 0.586, 0.914}*, при *k=10 – {0.042, 0.083, 0.133, 0.191, 0.259, 0.338, 0.447, 0.595, 0.901}*, при *k=12 – {0.035, 0.07, 0.106, 0.151, 0.202, 0.259, 0.323, 0.405, 0.514, 0.651, 0.958}*.

Кроме применения описанных методов дискретизации была также использована дискретизация на основе экспертных данных (интуитивная, Expert), были выбраны следующие точки разрыва: при *k=7 — {0.1, 0.2, 0.3, 0.5, 0.7, 1}*, при *k=9 — {0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 1, 1.5}*, при *k=10 — {0.05, 0.1, 0.2, 0.3, 0.4, 0.5 0.7, 1, 1.5}*, при *k=12 — {0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1, 1.5}*.

IV. Анализ работы модели при разных разбиениях случайной величины λ

Далее на полученных данных было проведено автоматическое обучение параметров байесовской сети доверия. То есть для каждой пары переменных, соединенных ребром, были вычислены условные вероятности. Таким образом, было получено 12 моделей, каждая обученная при различном разбиении случайной непрерывной величины λ, характеризующей интенсивность поведения.

Основная задача построенной модели – автоматизация оценки интенсивности при условии, когда доступны лишь сведения о величине интервалов между несколькими последними эпизодами, минимальном и максимальном значении величин интервалов между эпизодами. При дискретизации значений интенсивности эта задача представляет собой задачу классификации по *k* классам. Таким образом, одна из главных характеристик работы модели в данном случае является средняя точность (average accuracy), то есть доля правильно классифицированных значений в сумме матриц ошибок (confusion matrix) для каждого класса. Рассмотрим точность (accuracy, соотношение числа верно предсказанных значений, относительно всех значений) и среднюю точность предсказания модели при различных разбиениях величины λ. В таблице I показаны полученные результаты (при округлении до тысячных).

ТАБЛИЦА I Результаты предсказаний модели при различных дискретизациях λ

Метод дискретизации	*k*	Точность	Средняя точность
EW	7	0.653	0.901
EW	9	0.591	0.909
EW	10	0.578	0.916
EW	12	0.553	0.925
EF	7	0.516	0.862
EF	9	0.448	0.877
EF	10	0.408	0.882
EF	12	0.373	0.896
EF_Unique	7	0.531	0.867
EF_Unique	9	0.456	0.879
EF_Unique	10	0.413	0.883
EF_Unique	12	0.358	0.893
Expert	7	0.552	0.872
Expert	9	0.488	0.886
Expert	10	0.456	0.891
Expert	12	0.429	0.904

V. Выводы по результатам вычислительного эксперимента

Наиболее высокую точность показала модель, обученная при дискретизации алгоритмом EW при *k=7* (0.653), а самую высокую среднюю точность — модель, обученная при дискретизации алгоритмом EW при *k=12* (0.925). Самая низкая точность была получена при EF_Unique и *k=12* (0.358), а самая низкая средняя точность при EF_Unique и *k=7* (0.867). Разбиение алгоритмом EW в целом показало лучшие результаты, однако при использовании этого метода дискретизации и предсказаний модели на реальных данных можно получить ухудшение этих характеристик за счет большего числа резко отличающихся экземпляров процесса. Экспертное разбиение также показало хорошие результаты, это значит, что при применении модели в различных сферах этот шаг исследований может быть доверен эксперту в исследуемой области. Несмотря на то, что EF_Unique – это модифицированный EF, его применение не дало значительных положительных изменений относительно EF. Можно заметить, что с увеличением количества интервалов точность уменьшается, однако при этом средняя точность немного увеличивается, а в данном случае именно она является основной характеристикой результативности модели.

Отметим, что средняя точность во всех случаях довольно высока и варьируется в промежутке от 0.867 до 0.925, что говорит о возможности применять данную модель при различных дискретизациях величины, характеризующей интенсивность поведения.

VI. Заключение

В ходе исследования был осуществлён анализ работы модели оценки интенсивности поведения при использовании различных разбиений непрерывной

величины, характеризующей интенсивность процесса. Показано, что на синтетических данных модель показывает высокое качество работы при разных вариантах дискретизации.

Данная работа подготавливает базу для дальнейших исследований по совершенствованию модели оценки интенсивности поведения и апробации на реальных данных. При этом нужно обеспечить, чтобы исходная интенсивность поведения была известна. Такие данные могут быть получены, например, из социальных сетей, где многие действия пользователей, которые могут быть выбраны в качестве наблюдаемого поведения, фиксируются во времени.

В дальнейшем полученные результаты могут найти применение во многих областях, где необходимо оценить интенсивность некоторого поведения, имея при этом ограниченный сведениями о последних эпизодах поведения набор данных.

Список литературы

[1] Friman P.C. Cooper, Heron, and Heward's applied behavior analysis (2nd ed.): checkered flag for students and professors, yellow flag for the field // Applied Behavior Analysis. 2013. No. 1. P. 161–174. DOI: 10.1901/jaba.2010.43-161.

[2] Khlobystova A.O., Abramov M.V., Tulupyev A.L. Soft Estimates for Social Engineering Attack Propagation Probabilities Depending on Interaction Rates Among Instagram Users // International Symposium on Intelligent and Distributed Computing. Springer, Cham, 2019. P. 272–277. DOI: 10.1007/978-3-030-32258-8_32.

[3] Luhmann M. Using Big Data to study subjective well-being // Current Opinion in Behavioral Sciences. 2017. Vol. 18. P. 28–33. DOI: 10.1016/j.cobeha.2017.07.006.

[4] Wilson A.M., Reynolds K.A., Verhougstraete M.P., Canales R.A. Validation of a Stochastic Discrete Event Model Predicting Virus Concentration on Nurse Hands // Risk Analysis. 2019. 39(8):1812.

[5] Ramírez-Gallego S., García S., Mouriño Talín H., Martínez-Rego D., Bolón-Canedo V., Alonso-Betanzos A., Benítez J.M., Herrera F. Data discretization: taxonomy and big data challenge // Wiley Interdisciplinary Reviews. 2016. 6(1). P. 5–21. DOI: 10.1002/widm.1173.

[6] Couso I., Borgelt C., Hullermeier E., Kruse R. Fuzzy Sets in Data Analysis: From Statistical Foundations to Machine Learning // IEEE Computational Intelligence Magazine. 2019. Vol. 14, no. 1. P. 31–44. DOI: 10.1109/MCI.2018.2881642.

[7] Chen Y.-C.,Wheeler T., Kochenderfer M. Learning Discrete Bayesian Networks from Continuous Data // Journal of Artificial Intelligence Research. 2015. No. 59. DOI: 10.1613/jair.5371.

[8] Торопова А.В., Абрамов М.В., Тулупьева Т.В. Машинное обучение байесовской сети доверия как инструмента оценки интенсивности процесса по данным из социальной сети // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21. № 5. С. 727–737. Doi: 10.17586/2226-1494-2021-21-5-727-737.

[9] Toropova A.V., Tulupyeva T.V. Synthesis and learning of socially significant behavior model with hidden variables // Advances in Intelligent Systems and Computing. 2019. Vol. 875. P. 76–84.

[10] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. URL: http://www.R-project.org (дата обращения: 11.03.2022).

[11] bnlearn - an R package for Bayesian network learning and inference. URL: https://www.bnlearn.com/ (дата обращения: 11.03.2022).

[12] Jiang F., Sui Y. A novel approach for discretization of continuous attributes in rough set theory // Knowledge-Based Systems. 2015. 73. P. 324–334. DOI: 10.1016/j.knosys.2014.10.014.

[13] Hacibeyoglu M., Ibrahim M.H. EF_Unique: An Improved Version of Unsupervised Equal Frequency Discretization Method // Arabian Journal for Science and Engineering. 2018. 43. P. 7695–7704. DOI: 10.1007/s13369-018-3144-z.

[14] Cebeci Z., Yildiz F. Unsupervised discretization of continuous variables in a chicken egg quality traits dataset // Turkish Journal of Agriculture - Food Science and Technology. 2017. 5(4). P. 315–320. DOI: 10.24925/turjaf.v5i4.315-320.1056.

[15] Morente-Molinera J.A., Mezei J., Carlsson C., Herrera-Viedma E. Improving Supervised Learning Classification Methods Using Multigranular Linguistic Modeling and Fuzzy Entropy // IEEE Transactions on Fuzzy Systems. 2017. 25(5). P. 1078.