Применимость коэффициентов сходства в задаче сравнения социального окружения

А. А. Корепанова[1], В. Д. Олисеенко[2], М. В. Абрамов[3]
Санкт-Петербургский институт информатики и автоматизации Российской академии наук;
Санкт-Петербургский государственный университет
Санкт-Петербург, Россия
[1]aak@dscs.pro, [2]vdo@dscs.pro, [3]mva@dscs.pro

***Аннотация*. В данном исследовании представлен результат сравнения и выбора коэффициента сходства для сопоставления социального окружения пользователей в контексте решения задачи нахождения профилей одного и того же пользователя в различных социальных сетях. Выбранный коэффициент позволит более точно сравнивать социальное окружение пользователей и уточнить оценки бинарного классификатора (логистической регрессии) для определения профиля одного и того же пользователя в различных социальных сетях.**

Полученные результаты могут быть применены в задаче агрегации данных из социальных сетей о пользователе информационной системы и последующего построения профиля его уязвимостей, а также в других исследованиях, посвящённых социальным сетям или сравнению социального окружения.

Ключевые слова: коэффициенты сходства; социальные сети; идентификация пользователя; социоинженерные атаки; машинное обучение; информационная безопасность; защита пользователя; профиль уязвимостей пользователя

I. Введение

Вопросы информационной безопасности не теряют своей актуальности в последнее время [1]. В частности, эксперты отмечают возрастающую угрозу социоинженерных атак на информационные системы [2]. В связи с этим все более актуальными становятся исследования, посвящённые повышению защищённости информационных систем от атак этого типа. Для оценки защищённости пользователей информационных систем от социоинженерных атак удобно использовать профили уязвимостей пользователя [3] Одним из источников данных для построения фрагмента профиля уязвимостей пользователя являются социальные сети [7]. Для наиболее эффективного использования этого источника полезно было бы собирать данные из аккаунтов пользователя информационной системы в различных социальных сетях, так как при анализе большего числа аккаунтов можно извлечь больше потенциально полезных данных. Эти аккаунты не всегда могут быть известны, так что возникает проблема поиска аккаунтов пользователя в социальных сетях. В качестве этапа её решения, разрабатывается подход сопоставления различных аккаунтов и определения тех из них, что принадлежат одному пользователю [9].

В рамках предложенного подхода сопоставляются не профили напрямую, а составленные по ним мета-профили, которые представляются как набор значений унифицированных атрибутов. Мета-профили сопоставляются через вычисление коэффициентов сходства значений соответствующих атрибутов с помощью методов точного и нечёткого сравнения. Одним из этих атрибутов является социальное окружение пользователя, представленное списком друзей пользователя. В работе [9] для сопоставления значений этого атрибута использовался коэффициент Браун-Бланке, цель данного исследования состоит в том, чтобы проверить, можно ли улучшить сопоставление за счёт применения других мер сходства. Среди наиболее востребованных в аналогичных задачах коэффициентов сходства для анализа выбраны следующие: коэффициент Серенсена, коэффициент Кульчинского, коэффициент Отиаи, коэффициент Шимкевича–Симпсона, коэффициент Браун–Бланке, коэффициент Жаккара.

II. Постановка задачи

Наилучшим коэффициентом социального сходства будет считаться тот коэффициент, который улучшает оценку качества логистической регрессии для бинарной классификации со следующим условием: пусть X – множество пар профилей пользователей социальных сетей «ВКонтакте» и «Одноклассники», а Y – множество классов {0; 1}, где 0 означает, что пара профилей не принадлежит одному пользователю, а 1 – что принадлежит. Для сопоставления профилей используются мета-профили. Мета-профиль пользователя рассматривается как набор значений следующих атрибутов: «фамилия», «имя», «город», «возраст» и «список друзей», где «список друзей» представляется как набор мета-профилей «друзей» профиля. Признаками классификации выступают числовые характеристики подобия значений соответствующих атрибутов. Для вычисления численных коэффициентов подобия применяются как методы точного, так и различные методы нечёткого сравнения. Более подробное описание применяемых методов, программных инструментов и собранного датасета представлено в статье [9]. Данная работа сфокусирована на следующей задаче: необходимо сравнить несколько бинарных коэффициентов

Работа выполнена в рамках проекта по государственному заданию СПИИРАН № 0073-2019-0003, при финансовой поддержке РФФИ, проекты №18-01-00626 и №20-07-00839.

сходства для атрибута «список друзей» и определить наилучший для повышения точности классификации. Рассматриваемые коэффициенты:

- Коэффициент Жаккара: $K_J = \frac{c}{a+b-c}$,
- коэффициент Серенсена: $K_S = \frac{2c}{a+b}$,
- коэффициент Кульчинского: $K_K = \frac{c}{2}\left(\frac{1}{a}+\frac{1}{b}\right)$,
- коэффициент Отиаи: $K_O = \frac{c}{\sqrt{ab}}$,
- коэффициент Шимкевича: $K_{SZ} = \frac{c}{\min(a,b)}$,
- коэффициент Браун–Бланке: $K_B = \frac{c}{\max(a,b)}$,

где a – число друзей первого профиля, b – число друзей второго профиля, c – число совпавших друзей первого и второго профиля. Два друга считаются совпавшими, если мера сходства Джаро–Винклера между их именами и фамилиями больше 0.8.

III. Релевантные работы

Проблема выбора наилучшего коэффициента сходства в задаче нахождения аккаунтов одного и того же пользователя в различных социальных сетях не имеет единого, устоявшегося подхода для решения. Так, например, в статье [10], посвященной социальным сетям «Twitter» и «Facebook», используется коэффициент Сёренса для получения меры сходства социального круга пользователей. В работах [12] при решении аналогичной задачи используется коэффициент Жаккара. Именно поэтому применимость других коэффициентов сходства для решения предложенной задачи вызывает интерес.

Часто применение коэффициентов сходства находится в области биологии и медицины, например, одним из фундаментальных исследований применимости коэффициентов сходства является исследование, посвящённое проблеме формирования клеток [14], что подчеркивает важность применения данных коэффициентов в других областях.

IV. Определение наилучшего коэффициента

Для выбора наилучшего коэффициента можно использовать такие модели отбора, как метод последовательного отбора (Stepwise regression), метод прямого отбора (Forward selection), метод обратного отбора (Backward elimination), метод «лучших подмножеств» (Best Subsets) [15], которые в нашем случае основаны на логистической регрессии и тесте отношения правдоподобия. Суть данных методов заключается в поступательном отборе признаков на основе теста отношения правдоподобия. Однако, для их применения необходимым условием является отсутствие коллинеарности коэффициентов. Для проверки коллинеарности рассчитаем корреляцию между коэффициентами. Для выбора типа корреляции проверим коэффициенты на нормальность при помощи теста Шапиро–Уилка (табл. I).

ТАБЛИЦА I Тест Шапиро–Уилка

		Значение статистики	***p-value***
Название коэффициента	***Жаккара***	0.74	8.68e–26
	Серенсена	0.75	3.59e–25
	Кульчинского	0.52	4.50e–33
	Отиаи	0.79	1.13e–23
	Шимкевича	0.40	1.02e–35
	Браун-Бланке	0.72	1.34e–26

Полученные результаты показывают отсутствие нормального распределения у коэффициентов, таким образом в качестве коэффициента корреляции возьмём корреляцию Спирмена.

Рассмотрим подробнее результаты, представленные в табл. II. Некоторые коэффициенты имеют высокую корреляцию (>0.9) между собой. Например, коэффициенты Жаккара и Брауна–Бланке. Это означает, что отбирать наилучшие коэффициенты посредством предложенных методов будет некорректно. Поэтому подойдем к данному вопросу с двух точек зрения: первая основана на методах машинного обучения, вторая – на статистических методах.

ТАБЛИЦА II Корреляция коэффициентов

	K_B	K_S	K_J	K_K	K_O	K_{SZ}
K_B	1	0.73	0.96	0.81	0.89	0.96
K_S	0.73	1	0.77	0.93	0.84	0.77
K_J	0.96	0.77	1	0.84	0.93	1
K_K	0.81	0.93	0.84	1	0.91	0.84
K_O	0.89	0.84	0.93	0.91	1	0.93
K_{SZ}	0.96	0.77	1	0.84	0.93	1

Рассмотрим подход на основе машинного обучения. Построим модели логистической регрессии с признаками «ИФГ», «возраст», «коэффициент сходства», где «возраст» – числовая характеристика подобия значений соответствующего атрибута профилей, «ИФГ» – среднее арифметическое результатов сравнения значений атрибутов «имя», «фамилия», «город», а «коэффициент сходства» для каждой модели – один из сравниваемых коэффициентов, применённый к значениям атрибута «список друзей». У построенных моделей, используя процедуру 4fold cross-validation, сравним AUC (рис. 1), коэффициент детерминации Макфаддена $\text{McFadden R}^2 = 1 - \frac{log(L_c)}{log(L_{null})}$, где $log(L_c)$ – значение логарифма функции максимального правдоподобия рассчитанной модели, а $log(L_{null})$ – остаток логарифма

функции правдоподобия, и метрика $\text{Accuracy} = \frac{T}{N}$, где T – количество правильно классифицированных пар пользователей, а N – общее количество элементов (табл. III).

ТАБЛИЦА III СРАВНЕНИЕ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

		R^2	*Accuracy*
Название коэффициента	***Жаккара***	0.46	0.84
	Серенсена	0.56	0.90
	Кульчинского	0.61	0.91
	Отиаи	0.60	0.90
	Шимкевича	**0.64**	**0.92**
	Браун-Бланке	0.51	0.86

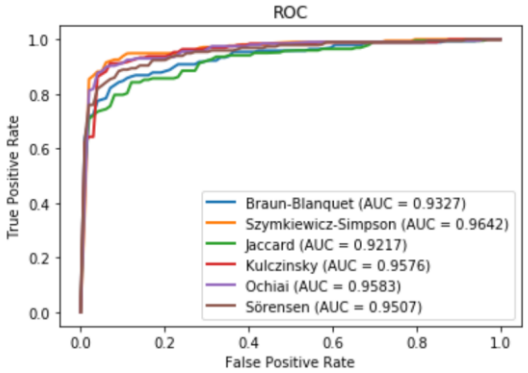

Рис. 1. Сравнение моделей

По полученным результатам можно сделать вывод, что коэффициент Шимкевича–Симпсона является наиболее предпочтительным коэффициентом, однако некоторые источники [16] утверждают, что сравнение при помощи AUC является ненадежным в отборе признаков при одинаковых типах моделей. Поэтому подтвердим полученные результаты при помощи статистических методов AIC и BIC (табл. IV).

ТАБЛИЦА IV AIC и BIC

		AIC	*BIC*
Название коэффициента	***Жаккара***	158.39	174.78
	Серенсена	157.75	174.14
	Кульчинского	146.85	167.23
	Отиаи	182.03	198.42
	Шимкевича	**132.63**	**143.01**
	Браун–Бланке	164.98	181.37

Таким образом, наилучшим коэффициентом для сравнения социального окружения является коэффициент Шимкевича–Симпсона.

V. ЗАКЛЮЧЕНИЕ

В данной работе было исследовано применение шести бинарных коэффициентов сходства – коэффициента Серенсена, коэффициента Кульчинского, коэффициента Отиаи, коэффициента Шимкевича–Симпсона, коэффициента Браун–Бланке, коэффициента Жаккара – к сопоставлению социального окружения профилей различных социальных сетей в рамках алгоритма определения принадлежности профилей одному пользователю. С применением каждого коэффициента были построены модели бинарной классификации. В результате анализа построенных моделей был выбран коэффициент Шимкевича–Симпсона. Данные результаты способствуют повышению точности алгоритма сопоставления профилей, предложенного в [9], который применяется в задаче составления профиля уязвимостей пользователя, за счет агрегации сведений о большем числе параметров. Кроме того, они могут быть использованы при анализе социальных сетей, например, в задачах социокомпьютинга. Возможные дальнейшие направления исследования состоят в увеличении эффективности модели классификации за счёт уточнения оценок сходства значений остальных атрибутов мета-профиля пользователя, а также выделении новых атрибутов.

СПИСОК ЛИТЕРАТУРЫ

[1] Исследование Cisco свидетельствует о росте инвестиций в технологии автоматизации и облачной информационной безопасности [Электронный ресурс]. URL: https://ru-bezh.ru/press-releases/33285-issledovanie-cisco-svidetelstvuet-o-roste-investiczij-v-texnolog (дата обращения: 23.03.2020)

[2] Social engineering attacks become more complex, advanced [Электронный ресурс]. URL: https://www.itweb.co.za/content/5yONP7Eg2QZqXWrb (дата обращения: 23.03.2020)

[3] Багрецов Г.И., Шиндарев Н.А., Абрамов М.В., Тулупьева Т.В. Подходы к автоматизации сбора, структурирования и анализа информации о сотрудниках компании на основе данных социальной сети // Нечеткие системы, мягкие вычисления и интеллектуальные технология (НСМВИТ-2017) труды VII всероссийской научной-практической конференции. 2017. С. 9-16.

[4] Khlobystova A.O., Abramov M.V., Tulupyev A.L. An approach to estimating of criticality of social engineering attacks traces // Studies in Systems, Decision and Control. P. 446–456.

[5] Khlobystova A.O., Abramov M.V., Tulupyev A.L. Identifying the most critical trajectory of the spread of a social engineering attack between two users // The Second International Scientific and Practical Conference “Fuzzy Technologies in the Industry – FTI 2018”. CEUR Workshop Proceedings. P. 38–43

[6] Азаров А.А., Тулупьева Т.В., Тулупьев А.Л. Прототип комплекса программ для анализа защищенности персонала информационных систем, построенный на основе фрагмента профиля уязвимостей пользователя // Труды СПИИРАН. 2012. № 2 (21). С. 21-40.

[7] Абрамов М.В., Тулупьев А.Л., Сулейманов А.А. Задачи анализа защищенности пользователей от социоинженерных атак: построение социального графа по сведениям из социальных сетей // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 2. С. 313-321.

[8] Kharitonov, N.A., Maximov, A.G., Tulupyev, A.L. Algebraic Bayesian Networks: Naïve Frequentist Approach to Local Machine Learning Based on Imperfect Information from Social Media and Expert Estimates // Communications in Computer and Information Science. 2019. P. 234-244

[9] Корепанова А.А., Олисеенко В.Д., Абрамов М.В., Тулупьев А.Л. Применение методов машинного обучения в задаче идентификации аккаунтов пользователя в двух социальных сетях // Компьютерные инструменты в образовании. 2019. №3. С. 29–43.

[10] Коршунов А., Белобородов И., Бузун Н., Аванесов В., Пастухов Р., Чихрадзе К., Козлов И., Гомзин А., Андрианов И., Сысоев А., Ипатов С., Филоненко И., Чуприна К., Турдаков Д., Кузнецов С. Анализ социальных сетей: методы и приложения // Труды ИСП РАН. 2014.

[11] Paridhi J., Ponnurangam K., Anupam J. @I seek 'fb.me': Identifying users across multiple online social networks // Conference: Proceedings of the 22nd international conference on World Wide Web companion.

[12] Waseem A. Rashid A. Social Account Matching in Online Social Media using Crosslinked Posts // International Conference on Pervasive Computing Advances and Applications – PerCAA 2019. P. 222–229.

[13] Bennacer N., Nana Jipmo C., Penta A., Quercini G. (2014) Matching User Profiles Across Social Networks. // Advanced Information Systems Engineering. CAiSE 2014. Lecture Notes in Computer Science. vol 8484.

[14] Yin Y., Yasuda K.. Similarity coefficient methods applied to the cell formation problem: A taxonomy and review // International Journal of Production Economic. Volume 101. Issue 2. June 2006. P. 329-352.

[15] Методы отбора переменных в регрессионные модели [Электронный ресурс]. URL: https://basegroup.ru/community/articles/feature-selection (дата обращения: 23.03.2020).

[16] Seshan V.E., Gonen M., Begg C.B. Comparing ROC Curves Derived From Regression Models // Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series. 2011. №20 URL: https://biostats.bepress.com/mskccbiostat/paper20 (дата обращения: 23.03.2020).