ВЕСТНИК ВОЕННОГО ИННОВАЦИОННОГО ТЕХНОПОЛИСА «ЭРА», 2022, том 3, № 2, с. 222–224

КОГНИТИВНЫЕ И СОЦИОГУМАНИТАРНЫЕ ИССЛЕДОВАНИЯ

УДК 612.821; 343.98

ПРИМЕНЕНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ ДЛЯ АНАЛИЗА ПОЛИГРАММ

© 2022 г. И. С. Лисицин[1], В. А. Орлов[1], Д. Г. Малахов[1], Л. И. Скитева[1,*]

[1]Национальный исследовательский центр "Курчатовский институт", Москва, Россия

******E-mail: skiteva_li@nrcki.ru*

Поступила в редакцию 01.04.2022 г.
После доработки 01.04.2022 г.
Принята к публикации 15.04.2022 г.

Описано программное обеспечение, основанное на реализации алгоритма анализа с применением методов кластеризации, предлагаемое к использованию при анализе полиграмм как дополнение к экспертной оценке. Разработан алгоритм, позволяющий проводить оценку полиграфических тестов без необходимости использования дополнительных данных для обучения, показана его эффективность в ходе исследований с применением полиграфа.

DOI: 10.56304/S2782375X22020085

ВВЕДЕНИЕ

Исследования с применением полиграфа (**ИПП**) активно используются как в оперативно-розыскной и судебной деятельности, так и в кадровом менеджменте предприятий [1]. Существуют различные алгоритмы, позволяющие в ручном, автоматическом или комбинированном режиме проводить анализ данных, полученных в ходе ИПП. Однако все они имеют ряд критических недостатков.

Широко применяемые в настоящее время методы экспертной оценки полиграмм характеризуются трудоемкостью и значительными временными затратами [2, 3], кроме того, они не избавлены от субъективизма – достоверность заключения таких методов зависит от профессиональной квалификации и опыта полиграфолога.

Также для анализа полиграмм применяется метод логистической регрессии [4], требующий проведения нескольких контрольных тестов для подсчета весовых коэффициентов.

Таким образом, необходима разработка алгоритма анализа полученных в ходе ИПП данных, который исключал бы зависимость заключения от человеческого фактора полиграфолога, не требуя при этом большого объема данных.

Одним из перспективных методов обработки полученных в ИПП данных является применение алгоритмов машинного обучения. Однако разработанный в [5] персептрон для анализа требует обучения на относительно большой выборке данных (для каждого обследуемого человека формируется выборка из ~70–100 вопросов–ответов для обучения персептрона), что крайне трудно реализовать в рамках ИПП. Чтобы исключить описанный недостаток, в настоящем исследовании для решения поставленной задачи использовали методы кластеризации, не требующие для обучения дополнительных данных [6].

МЕТОДЫ

Исследование проводили на данных, полученных в результате прохождения добровольцами в рамках ИПП теста на скрываемое имя (**ТСИ**). При проведении теста использовали пять нейтральных стимулов и один значимый. Реакции на один из нейтральных стимулов, первый по счету предъявлений (нулевое предъявление), не учитывали, так как он считается адаптационным. В ТСИ предъявлениями служат названные проводящим исследование экспертом имена, одно из которых принадлежит исследуемому человеку. В процессе проведения теста эксперт спрашивает, принадлежит ли названное имя исследуемому человеку, на что тот должен каждый раз отвечать отрицательно. Таким образом, известен стимул, на котором человек соврет – стимул с его именем и является значимым в данном тесте, на нем организм испытуемого покажет реакцию, отличную от фонового состояния, что и отразится на полиграмме. Каждый стимул в одном ТСИ предъявляется по 5 раз.

Выбор такого теста для исследования обусловлен возможностью оценить точность заключений разработанного алгоритма, так как в нем известен

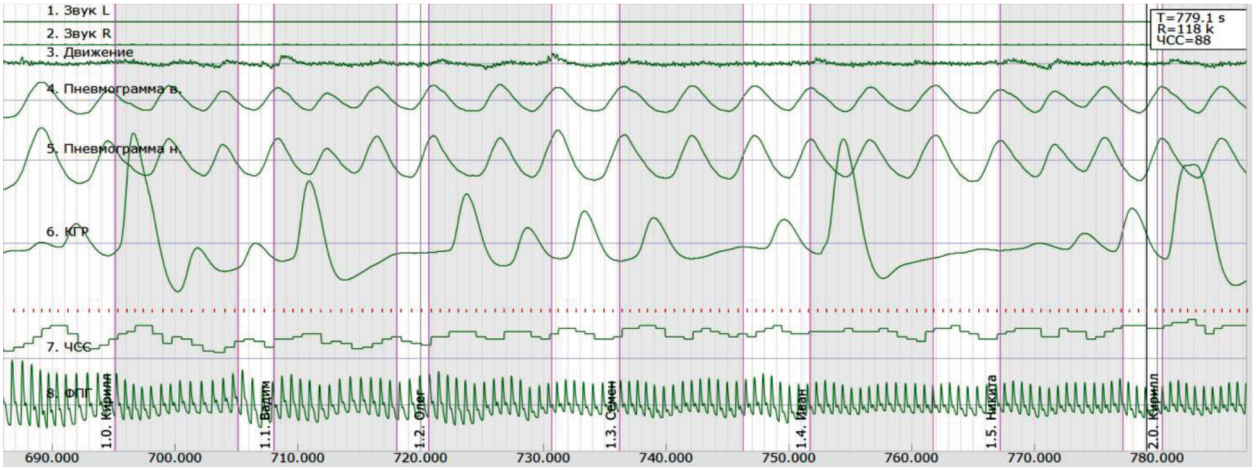

Рис. 1. Фрагмент полиграммы испытуемого при прохождении теста на скрываемое имя.

значимый стимул. Всего в исследовании использовали 122 теста.

На рис. 1 представлен фрагмент полиграммы, записанной при прохождении испытуемым ТСИ. Для анализа использовали данные, полученные с пяти каналов: пневмограммы (верхней и нижней), кожно-гальванической реакции, частоты сердечных сокращений и фотоплетизмограммы.

Разработанное в настоящем исследовании программное обеспечение проводит анализ на основе записанной в ИПП полиграммы. На вход алгоритм получает длину линии каждого из пяти исследуемых каналов в течение 10 с после предъявления стимула.

Кластеризацию входных данных проводили методами k-средних и иерархической кластеризации.

При разделении выборки на кластеры использовали деление на два и на три кластера. При делении на два кластера результаты разделяли на фоновые реакции, которые чаще остальных встречаются на полиграмме, и отличные от них (реакции на значимый стимул или артефакты). Таким образом, наименьший по объему кластер будет содержать значимые реакции.

При делении на три кластера в отдельный кластер выделяли реакции, которые еще больше отличаются от остальных – к ним могут относиться либо так называемые артефакты (выбросы), либо сильные реакции на значимый стимул. Реализовано деление на три кластера, поскольку на полиграмме могут встретиться артефакты, выделяющиеся сильнее реакций на значимые стимулы. Если они имеются на полиграмме, то они будут выделены в кластер наименьшего объема и анализироваться будут так же, как второй по объему кластер.

Среди всех предъявлений, отнесенных алгоритмом к значимому кластеру, выбирается самый часто встречаемый стимул (так как предполагается, что артефакты, отнесенные к значимому кластеру, встречаются не закономерно, а случайно) – такой подход реализован для того, чтобы уменьшить влияние артефактов и неверно отнесенных к значимому кластеру реакций на итоговое заключение. Рассчитывали итоговое заключение из промежуточных результатов каждого метода кластеризации по формуле Байеса:

$$P(H_i \mid A) = \frac{P(H_i)*P(A \mid H_i)}{\sum_{k=1}^{n} P(H_k)*P(A \mid H_k)}, \qquad (1)$$

где вероятности каждого из пяти стимулов выделения в качестве значимого образуют полную группу, $P(H_k)$ – вероятность, с которой среди промежуточных заключений встретится k-стимул, $P(A \mid H_k)$ – вероятность, с которой k-стимул встретится среди всех выделенных методами кластеризации стимулов.

Методом k-средних алгоритмом был верно определен значимый стимул в 74.6% тестов при делении на два кластера и в 79.5% при делении на три кластера. Аналогичные результаты получены при использовании метода иерархической кластеризации (70.5% верных заключений при делении на два кластера и 76.2% при делении на три кластера). Деление начальной выборки на более чем три кластера существенно уменьшает процент верного заключения (<65% верных заключений). Однако результаты при делении на два и на три кластера часто, но незначительно различались. Было решено делать общее заключение, основываясь на результатах двух вариантов кластеризации.

Метод *k*-средних минимизирует суммарные квадратичные отклонения точек кластеров от центроидов этих кластеров. Он случайным образом относит каждое наблюдение к одному из кластеров, а затем переназначает наблюдения для минимизации евклидовых расстояний между каждым наблюдением и центроидом.

Алгоритм иерархической кластеризации начинает работу с того, что каждому экземпляру данных сопоставляется свой собственный кластер. Затем два ближайших кластера объединяются в один и так далее, пока не будет образован один общий кластер. Для деления исходной выборки применяется кластеризация на основе деревьев.

Описываемые методы кластеризации имеют разную временную сложность: линейная для метода *k*-средних и квадратичная для метода иерархической кластеризации.

В кластеризации при помощи метода *k*-средних алгоритм начинает определение центроида со случайных точек в пространстве, поэтому результаты, генерируемые при многократном запуске алгоритма, могут различаться. Для метода иерархической кластеризации результаты воспроизводимы. Из-за описанных особенностей методов их результаты для одинаковых данных могут различаться.

При использовании метода *k*-средних было верно определено больше значимых стимулов (80.2%), но с меньшей их средней вероятностью (0.76). Метод иерархической кластеризации, наоборот, показал большую вероятность правильных заключений (0.85), сделав при этом меньше верных заключений (78.7%). Было решено использовать комбинацию двух методов, и эмпирический подход показал, что оптимальной их комбинацией является следующий алгоритм: сначала все данные обрабатываются методом *k*-средних. Если результат неоднозначный (в качестве значимых выделяются больше одного стимула), применяется метод иерархической кластеризации при делении на три кластера, так как такой метод показал наибольшую среднюю вероятность верных заключений. Полученный алгоритм позволяет верно определить значимый стимул в 83.4% тестов со средней вероятностью правильного заключения 0.83.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Полученные в ходе настоящего исследования результаты демонстрируют ранее не реализованную возможность применения методов кластеризации для решения задач, связанных с анализом полиграмм в оперативно-розыскной деятельности [1], для судебно-психофизиологических экспертиз, а также для задач клинической психиатрии [7, 8].

Из-за отсутствия объективных метрик качества эффективность анализа методами кластеризации трудно сравнивать с эффективностью применяемых в настоящее время алгоритмов оценки полиграмм. Однако часть обработанных в исследовании тестов были также обработаны с помощью метода логистической регрессии, среди них все заключения о выделении значимого стимула совпали для двух методов.

ЗАКЛЮЧЕНИЕ

Разработан алгоритм автоматического анализа полиграмм на основе методов кластеризации, позволяющий обрабатывать данные, полученные в ходе ИПП. Алгоритм исключает недостатки применяемых в настоящее время методов, так как позволяет проводить анализ отдельных тестов индивидуально, не требуя для обучения дополнительных данных.

СПИСОК ЛИТЕРАТУРЫ

1. *Холодный Ю.И.* Криминалистика: учебник для студентов вузов, обучающихся по направлению подготовки “Юриспруденция” / Под ред Бастрыкина А.И. и др. 3-е изд., перераб. и доп. М: ЮНИТИ-ДАНА: Закон и право, 2017. 799 с.
2. *Оглоблин С.И., Молчанов А.Ю.* Инструментальная “детекция лжи”: Академический курс. Ярославль: Нюанс, 2004.
3. *Ясницкий Л.Н.* Интеллектуальные информационные технологии и системы. Пермь: Изд-во Перм. ун-та, 2007.
4. *Леонтьев К.А., Панин С.Д., Холодный Ю.И.* // МГТУ Наука и Образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2014. № 10. С. 230.
5. *Ясницкий Л.Н., Петров А.М., Сичинава З.И.* // Изв. вузов. Поволжский регион. Технические науки. 2010. Т. 13. № 1. С. 64.
6. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning, Data Mining, Inference, and Prediction. New York: Springer, 2001. 533 p.
7. *Каменсков М.Ю.* // Российский психиатрический журнал. 2012. № 5. С. 14
8. *Захарова Н.В., Ковальчук М.В., Костюк Г.П. и др.* // Психическое здоровье. 2019. № 12. С. 50.