

ПРОВЕРКА ГИПОТЕЗЫ СОЗДАНИЯ ЦИФРОВОГО ПОЛИГРАФА НА ОСНОВЕ ВИДЕО И АУДИО ДАННЫХ

TESTING THE HYPOTHESIS OF CREATING A DIGITAL POLYGRAPH BASED ON VIDEO AND AUDIO DATA

В данной статье проводится проверка гипотезы возможности создания цифрового полиграфа на основе видео и аудио информации. Суть гипотезы заключается в том, что с развитием технологий, особенно информационных, появляется возможность за счет обработки видео и аудио информации обнаружить трансляцию правды/лжи, исходящую от человека. Предложенное решение может применяться для создания скрытых полиграфов, которые будут незаметны для испытуемых. Скрытые полиграфы позволят руководителям кадровых служб и служб безопасности судоходных компаний, портов и других объектов транспортной инфраструктуры убедиться в достоверности информации, предоставляемой сотрудниками, ответственными за транспортную безопасность. В данной статье описывается эксперимент получения видео и аудио данных с заведомо известной трансляцией правды/лжи. На основе кластерного анализа были выявлены информативные и неинформативные признаки, которые могут быть исключены из цифрового полиграфа как ненужные. Также были выявлены на основе анализа аудио-видео временные промежутки трансляции правды/лжи.

In this paper carried out to test the hypothesis the possibility of creating digital printing based on video and audio information. The essence of the hypothesis is that the development of technologies, especially information, it is possible due to the processing of video and audio data to detect broadcast of truth or lies emanating from the man. The proposed solution can be used to create a hidden polygraphs, which will be invisible to the subjects. Hidden polygraphs allow personnel managers and security of shipping companies, ports and other transport infrastructure to ensure the reliability of information provided to staff responsible for road safety. This article describes the experiment data acquisition video and audio data with a known translation of the well-known truth or lies. On the basis of cluster analysis identified informative and non-informative features that can be excluded from the digital polygraph as unnecessary. Just been identified based on the analysis of audio-video broadcast time intervals of truth or lies.

Ключевые слова: полиграф, кластерный анализ, видео, аудио.

Key words: polygraph, cluster analysis, video, audio.

Введение

Международный кодекс по охране судов и портовых средств (Кодекс ОСПС) предъявляет ряд требований к должностным лицам компаний и лицам командного состава, ответственных за охрану судов и портовых средств. Данные лица должны обладать достаточными знаниями и способностями для выполнения возложенных на них обязанностей [1].

Федеральным законом от 09.02.2007 г. №16-ФЗ «О транспортной безопасности» поставлена задача подготовки и аттестации сил обеспечения транспортной безопасности. Ст. 10 Закона запрещает выполнение работ, непосредственно связанных с обеспечением транспортной безопасности, лицам, имеющим непогашенные или неснятые судимости, состоящим на учете по поводу психических заболеваний, имеющим медицинские противопоказания, причастным к экстремистской деятельности и терроризму, сообщившим заведомо ложные сведения о себе при приеме на работу [2].

Опасность проникновения таких лиц в структуры обеспечения транспортной безопасности очевидна. Современные методы проверки персонала при приеме на работу, используемые кадровыми службами, не всегда позволяют выяснить степень достоверности предоставляемых работником данных о себе. Поэтому создание условий для скрытого контроля, правдивости предоставляемой на собеседовании информации от лиц, принимаемых на должности, связанные с вопросами транспортной безопасности, является актуальной задачей.

С развитием технологий, особенно информационных, появляется возможность создания цифровых полиграфов, использующих такую информацию как видео и аудио запись. Как известно, полиграф – это некое устройство (техническое), направленное на регистрацию психофизиологических реакций организма человека на трансляцию правды/лжи.

Целью работы была проверка гипотезы о возможности создания цифрового полиграфа, работающего на основе аудио и видео данных. В ходе эксперимента, были поставлены следующие задачи:

- 1) создание базы данных экспериментов с набором различных признаков трансляции правды/лжи;
- 2) проведение обезличивания базы данных с последующей ее подготовкой для проведения статистического анализа;
- 3) выбор подходящего метода проведения статистического анализа из числа возможных;
- 4) на основе статистических данных определение информативных и не информативных признаков;
- 5) определение временных интервалов трансляции правды/лжи.

Создание базы данных производилось путем записи на видео со звуком разговора с человеком на различные темы. Во время этих разговоров человек в определенный момент времени начинал лгать. После записи видео происходила его обработка и дальнейшее преобразование в набор цифровых данных в формат MS Excel (xls). Этих данных оказалось достаточно для проведения статистического анализа [3].

На основе поставленных задач и полученных данных был проведен краткий обзор основных методов статистического анализа, в число которых входят: корреляционный анализ, регрессионный анализ, канонический анализ, методы сравнения средних, частотный анализ, кросстабуляция, анализ соответствий, кластерный анализ, дискриминантный анализ, факторный анализ, деревья классификации, анализ главных компонент и классификация, многомерное шкалирование, моделирование структурными уравнениями, методы анализа выживаемости, временные ряды, нейронные сети [4].

Из всех указанных методов наиболее подходящими для решения поставленной задачи являются кластерный анализ и факторный анализ. Оба метода являются разделами многомерного статистического анализа.

Главными целями факторного анализа являются: сокращение числа переменных (редукция данных) и определение структуры взаимосвязей между переменными, т.е. классификация переменных. Поэтому факторный анализ используется или как метод сокращения данных, или как метод классификации.

Главной целью кластерного анализа является группировка многомерных объектов, основанная на представлении результатов отдельных наблюдений точками подходящего геометрического пространства с последующим выделением групп как сгустков этих точек (кластеров, таксонов).

Кластерный анализ реализует задачу сведения множества характеристик к небольшому ряду обобщающих итогов, выражающих действительно существенные для явления различия, позволяя при этом избежать отсека большинства признаков и возвращения к классической мало-размерной задаче. Именно поэтому кластерный анализ был выбран как основной метод реализации поставленной задачи.

Математическая основа кластерного анализа N измерений $X_1, X_2, \dots, X_n$ могут быть представлены в виде матрицы

$$X = [X_1, X_2, \dots, X_n] = \begin{bmatrix} x_{11} & x_{12} & x_{1n} \\ x_{21} & x_{22} & x_{2n} \\ \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{nn} \end{bmatrix} \quad (1)$$

Аналогичным образом расстояния между парами векторов $d(X_i, X_j)$ могут быть представлены в виде матрицы расстояний

$$\Delta = \begin{bmatrix} 0 & d_{12} & d_{1n} \\ d_{21} & 0 & d_{2n} \\ \dots & \dots & \dots \\ d_{n1} & d_{n2} & 0 \end{bmatrix}, \quad (2)$$

$d_{ii} = 0$, для $i = 1, 2, \dots, n$.

Если признаки измерены в разных единицах, то определить расстояние между объектами нельзя. Тогда применяется нормировка показателей, переводящая их в безразмерные величины. Наиболее распространенные способы нормирования следующие:

$$z1 = \frac{X - \bar{X}}{\sigma}; \quad z2 = \frac{X}{\bar{X}}; \quad z3 = \frac{X}{X_{\max}}; \quad z4 = \frac{X - \bar{X}}{X_{\max} - X_{\min}}. \quad (3)$$

Понятием, противоположным понятию расстояния между объектами X_i и X_j , является понятие близости (сходства) между X_i и X_j . Точнее, мера близости между объектами X_i и X_j – это вещественная функция $\mu(X_i, X_j) = \mu_{ij}$ со свойствами:

$$\begin{aligned} 0 \leq \mu(X_i, X_j) < 1 \text{ для } X_i \neq X_j; \\ \mu(X_i, X_i) &= 1; \\ \mu(X_i, X_j) &= \mu(X_j, X_i). \end{aligned} \quad (4)$$

Пары значений мер близости можно объединить в матрицу близости

$$\mu = \begin{bmatrix} 1 & \mu_{12} & \mu_{1n} \\ \mu_{21} & 1 & \mu_{2n} \\ \dots & \dots & \dots \\ \mu_{n1} & \mu_{n2} & 1 \end{bmatrix}, \quad (5)$$

$\mu_{ij} = 0$ для $i = 1, 2, \dots, n$.

Величину μ_{ij} называют коэффициентом близости. Примером линейной близости является коэффициент корреляции. Рассмотрим основные способы определения расстояний между объектами.

Метрики для количественных шкал (расстояние)

а) линейное расстояние

$$d(X_j, X_i) = \sum_{k=1}^N |x_{ki} - x_{kj}|; \quad (6)$$

б) Евклидово расстояние

$$d(X_j, X_i) = \left[\sum_{k=1}^N (x_{ki} - x_{kj})^2 \right]^{1/2}; \quad (7)$$

в) обобщенное степенное расстояние Минковского (универсальная метрика)

$$d(X_j, X_i) = \left[\sum_{k=1}^N (x_{ki} - x_{kj})^p \right]^{1/p}. \quad (8)$$

Метрики для качественных шкал (мера близости)

К качественным шкалам относят:

- а) номинальную шкалу (или шкалу наименований). Примеры измерения: пол (мужчина, женщина), национальность (француз, итальянец, немец), профессия (учитель, врач, бухгалтер) и др.;
- б) порядковую шкалу (или ранговую, ординарную). Примеры измерения: экспертные ранжировки, оценки предпочтений, шкала твердости минералов и др.

Расстояние для номинальных шкал вводится следующим образом. Пусть имеются два объекта X и Y с N признаками. Введем координаты x_i и y_j ($i = 1, 2, \dots, N$) как логические переменные, принимающие значение 1, если объект обладает 1-м признаком, и 0, если признак с номером 1 у объекта отсутствует.

Рассмотрим возможные способы точного определения кластеров:

- Класс типа сгущения (класс типа ядра): все расстояния между объектами внутри класса меньше любого расстояния между объектами класса и остальной частью множества.
- Класс с центром: класс называется классом с центром, если существует порог и некоторая точка x_i^* в пространстве, занимаемом объектами кластера S_1 со свойствами:

$$\tau > 0; \text{ если } d_{ix^*} \leq \tau, \text{ то } x_i \in S_1; \text{ если } d_{ix^*} > \tau, \text{ то } x_i \notin S_1. \quad (9)$$

Точка x_i^* называется *центром класса*. Часто в качестве x_i^* рассматривается центр тяжести, т.е. координаты центра определяются как средние значения признаков у объектов класса.

Далее, пусть $X = \{X_1, X_2, \dots, X_{n_1}\}$ – множество измерений, произведенных над множеством объектов $I = \{I_1, I_2, \dots, I_{n_2}\}$, а $Y = \{Y_1, Y_2, \dots, Y_{n_2}\}$ – множество измерений, соответствующее множеству $J = \{J_1, J_2, \dots, J_{n_2}\}$.

Величину $D = \frac{n_1 n_2}{n_1 + n_2} (\bar{X} - \bar{Y})^T (\bar{X} - \bar{Y})$, где $\bar{X} = \sum_{i=1}^{n_1} \frac{X_i}{n_1}$ и $\bar{Y} = \sum_{i=1}^{n_2} \frac{Y_i}{n_2}$, называют *расстоянием*

между кластерами I и J . Формула вычисления расстояния между кластерами используется как один из параметров в алгоритмах кластерного анализа.

На первом шаге, когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Далее необходимо правило объединения или связи для двух кластеров. Здесь имеются различные методы, в число которых входит метод k -средних.

Использование данного метода предполагает, что исследователь имеет гипотезу относительно числа кластеров (по наблюдениям или по переменным). В таком случае создается ровно k различных кластеров, расположенных на возможно больших расстояниях друг от друга. В качестве меры близости используется Евклидово расстояние:

$$p(x, y) = \|x - y\| = \sqrt{\sum_{p=1}^n (X_p - Y_p)^2}, \quad (10)$$

где $x, y \in R^n$.

Итак, рассмотрим ряд наблюдений $(x^{(1)}, x^{(2)}, \dots, x^{(m)})$, $x^{(j)} \in R^n$.

Метод k -средних разделяет m наблюдений на k групп (или кластеров)

$$(k \leq m), S = (S_1, S_2, \dots, S_k),$$

чтобы минимизировать суммарное квадратичное отклонение точек кластеров от центроидов этих кластеров:

$$\min \left[\sum_{i=1}^k \sum_{x^{(j)} \in S_i} \|x^{(i)} - \mu_i\|^2 \right], \quad (11)$$

где $x^{(i)} \in R^n$, $\mu_i \in R^n$; μ_i – центроид для кластера S_i .

Алгоритм

Итак, если мера близости до центроида определена, то разбиение объектов на кластеры сводится к определению центроидов этих кластеров. Число кластеров k задается исследователем заранее.

Рассмотрим первоначальный набор k средних (центроидов) $\mu_1, \dots, \mu_k$ в кластерах $S_1, \dots, S_k$. На первом этапе центроиды кластеров выбираются случайно или по определенному правилу (например, можно выбрать центроиды, максимизирующие начальные расстояния между кластерами).

Относим наблюдения к тем кластерам, чье среднее (центроид) к ним ближе всего. Каждое наблюдение принадлежит только к одному кластеру, даже если его можно отнести к двум и более кластерам. Затем центроид каждого i -го кластера перевычисляется по следующему правилу:

$$\mu_i = \frac{1}{s_j} \sum_{x^{(j)} \in S_i} x^{(j)}. \quad (12)$$

Таким образом, алгоритм k -средних заключается в перевычислении на каждом шаге центроида для каждого кластера, полученного на предыдущем шаге. Алгоритм останавливается, когда значения μ_i не меняются: $\mu_i^{\text{шаг } t} = \mu_i^{\text{шаг } t+1}$.

Важно отметить, что неправильный выбор первоначального числа кластеров k может привести к некорректным результатам. Именно поэтому при использовании метода k -средних важно сначала провести проверку подходящего числа кластеров для данного набора данных.

С применением данного метода была обработана полученная база данных, состоящая примерно из 50 экспериментов. На основе обработанных данных:

- получены информативные и неинформативные признаки, определены временные интервалы трансляции правды/лжи;
- выбраны двух и четырехкластерный анализ как наиболее подходящие для решения поставленной задачи (рис. 1, 2).

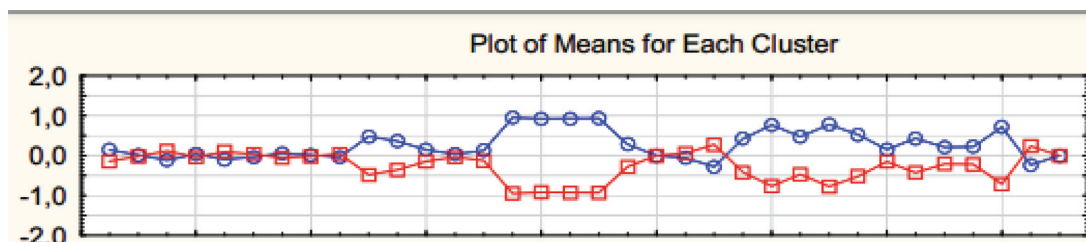


Рис. 1. Пример двухкластерного анализа для одного из экспериментов

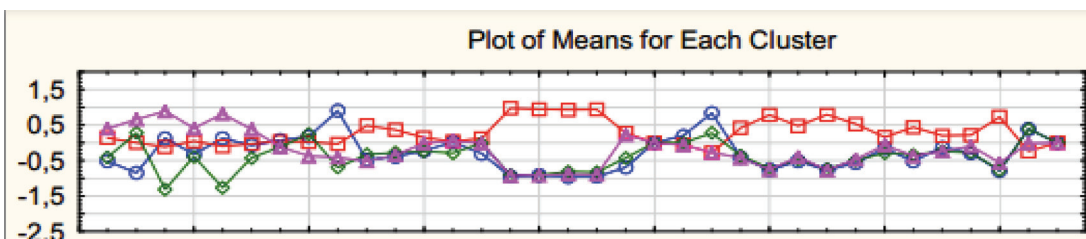


Рис. 2. Пример четырехкластерного анализа для одного из экспериментов

Для каждого исследования кластеры формировались отдельно. И, поскольку в кластере сохраняются моменты времени, то во всех случаях они разные, так как каждый испытуемый выдавал ложную/правдивую информацию разное количество времени.

Диаграмма, представленная ниже, основана на анализе полученных данных (49 файлов) методом кластерного анализа, она отображает степень информативности признаков (рис. 3).

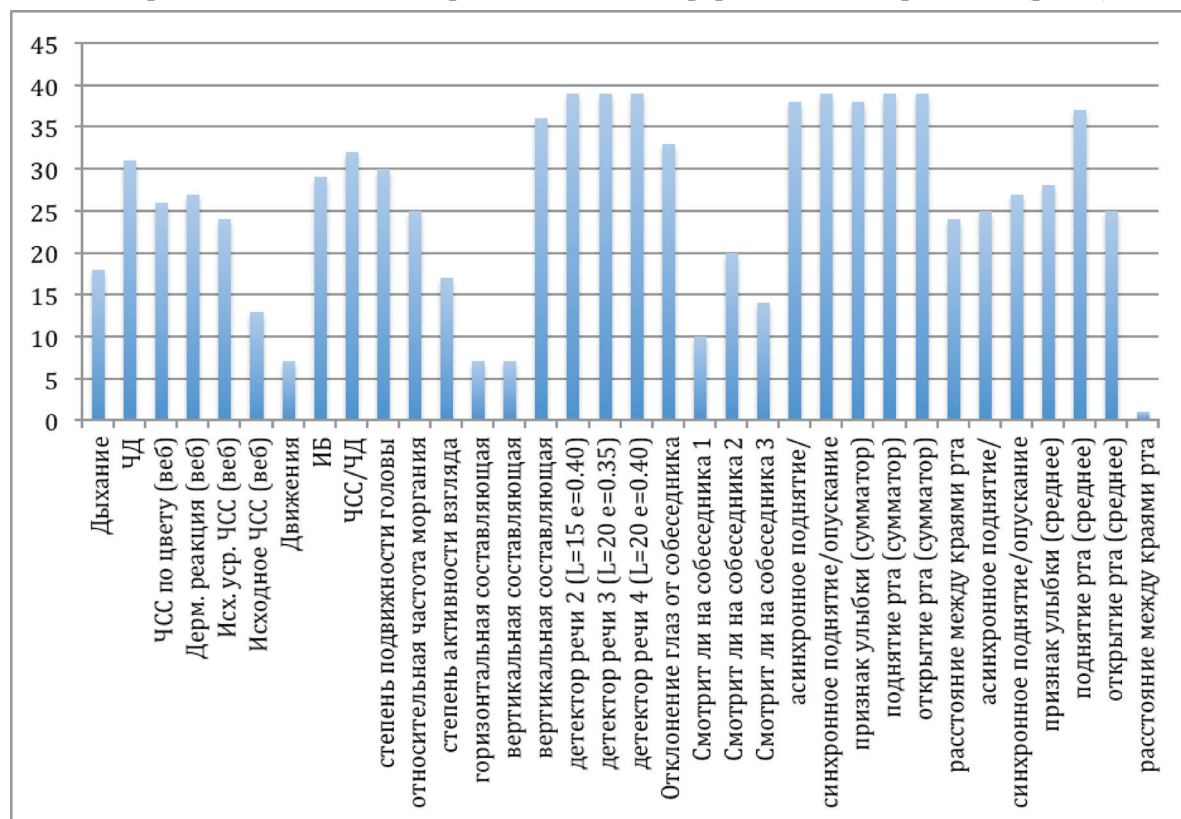


Рис. 3. Анализ степени информативности признаков

Из данной диаграммы видно, что такие признаки как частота дыхания (ЧД), частота сердечных сокращений (ЧСС), дерматическая реакция, индекс напряжения Р. М. Баевского (ИБ), ЧСС/ЧД, степень подвижности головы, вертикальная составляющая, детекторы речи, отклонение глаз от собеседника, асинхронное/синхронное поднятие/опускание бровей, признак улыбки, поднятие рта, открытие рта, поднятие рта (среднее) являются наиболее информативными. А такие признаки как исходное ЧСС, движения, горизонтальная и вертикальная составляющая направления взгляда, смотрит ли на собеседника и расстояние между краями рта практически не дают нам никакой полезной информации. Эта информация может оказаться полезной при оптимизации количества параметров для обработки. На самые информативные признаки можно обратить больше внимания, а неинформативные — убрать. Данная статистика была составлена после обработки базы данных экспериментов. На примерах двух- и четырехкластерного анализа отчетливо видно, что в некоторых моментах все кластеры сходятся в одном месте — это говорит нам о том, что на данном признаке показатели изменяются незначительно [5] – [14].

Далее, для анализа содержимого кластеров и для выявления тем 2–3 (спокойное состояние) и 22–23 (идет трансляция лжи) была написана программа на языке bash, которая выставляет время из двух кластеров (двухкластерный анализ) в хронологическом порядке, подсвечивая его разными цветами. Также на каждом моменте времени подписано, о какой теме идет речь. Из полученных данных видно, что практически все моменты времени, когда испытуемый говорил на темы 2–3 находятся в красном кластере (испытуемый был спокоен), а во время тем 22–23 моменты времени постоянно скачут из красного кластера в зеленый (т.е. испытуемый волнуется) (рис. 4).

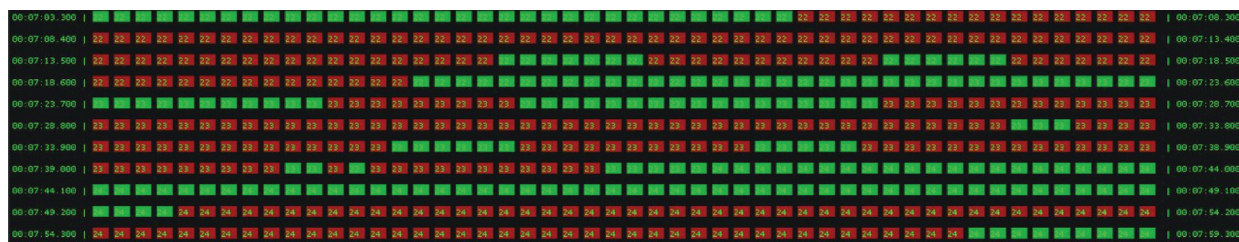


Рис. 4. Пример изображения, отображающего хронологию времени трансляции правды/лжи

Эти данные подтверждает график, полученный в программе Excel (рис. 5).

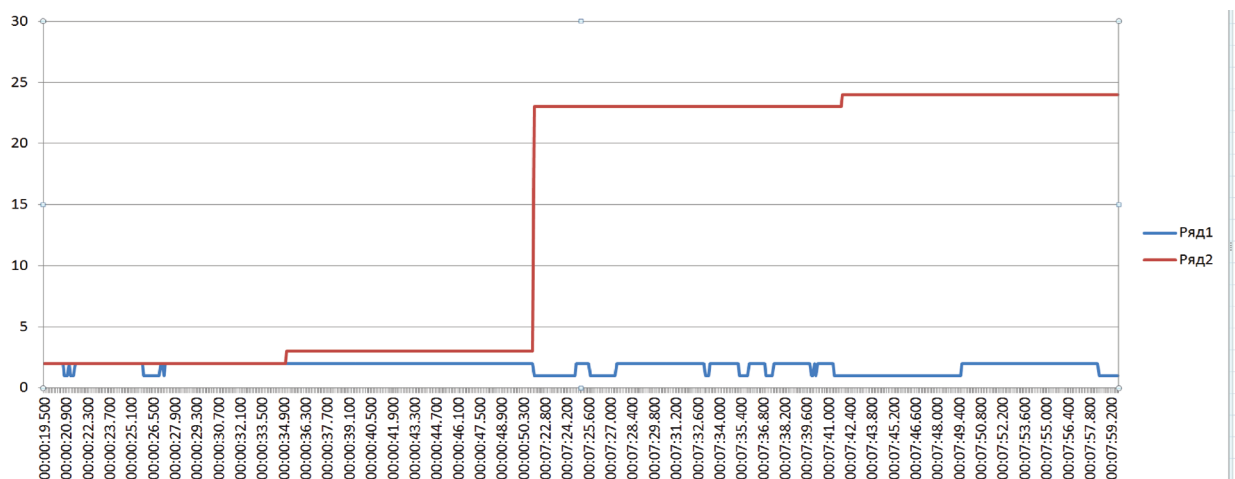


Рис. 5. График хронологии времени трансляции правды/лжи

Красный график — это темы. Здесь были выделены темы 2–3 и 22–23 для лучшей наглядности. Снизу показано время. А синий график — это кластеры: первый и второй. Хорошо видно, что пока испытуемый рассуждает на темы 2 и 3, колебание в кластерах минимально, но когда он переходит к темам 22–23, так же как и в программе, показанной ранее, перепад в кластерах намного заметнее.

Заключение

Из изложенного можно сделать вывод о том, что существует реальная возможность создания цифрового полиграфа на основе видео и аудио данных. Стоит отметить, что в каждом эксперименте испытуемый при трансляции заведомо ложных сведений переживает микростресс, что отражается на его вегетативной нервной системе и позволяет исследователю зафиксировать изменения. Изменения показателей закономерны, следовательно, к ним применимы методы статистического анализа.

С применением метода кластерного анализа исходные данные были разбиты по таксонам, отражающим два (или более, на усмотрение исследователя) состояния испытуемого. Уместно было бы соотносить эти состояния с трансляцией испытуемым заведомо ложных сведений. Также была создана статистика, отражающая информативность каждого отдельно взятого признака. На основе этой статистики можно соотнести информативность признаков между собой.

Учитывая современную возможность быстрой передачи больших массивов данных, коими являются видео и аудио данные, на любые расстояния, появляется реальная возможность в режиме практически реального времени подтверждать при помощи полиграфов истинность получаемой от субъектов информации, в том числе от субъектов, находящихся в рейсах и на удаленных объектах.

Список литературы

1. International Ship and Port Facility Security (ISPS) Code, 2nd edition, 2009. — 272 с.
2. Федеральный закон от 09.02.2007 №16-ФЗ (ред. от 03.02.2014) «О транспортной безопасности». [Электронный ресурс]. — Режим доступа: <http://base.garant.ru/12151931/#help> (дата обращения 13.05.2015).
3. Дронов С. В. Многомерный статистический анализ / С. В. Дронов. — Барнаул, 2005. — 213 с.
4. Гитис Л. Х. Статистическая классификация и кластерный анализ / Л. Х. Гитис. — М.: Изд-во Московского государственного горного ун-та, 2003. — 159 с.
5. Артемьев А. Б. Полиграф и права человека / А. Б. Артемьев, М. О. Воробьев // Юридическая мысль. — 2012. — № 2. — С. 13–18.
6. Семенцов В. А. Проверка правдивости показаний на полиграфе / В. А. Семенцов // Российский юридический журнал. — 2011. — № 2. — С. 111–116.
7. Леонтьев К. А. Оценка результатов тестирования на полиграфе методами регрессионного анализа / К. А. Леонтьев, С. Д. Панин, Ю. И. Холодный // Наука и образование: электронное научно-техническое издание. — 2014. — № 10. — С. 230–243.
8. Гаужаева В. А. Методические основы проведения исследований на полиграфе / В. А. Гаужаева // Теория и практика общественного развития. — 2015. — № 1. — С. 57–60.
9. Савченко Т. Н. Применение методов кластерного анализа для анализа данных психологических исследований / Т. Н. Савченко // Прикладная юридическая психология. — 2008. — № 4. — С. 100–111.
10. Нелипа Д. В. Применение кластерного анализа при проведении системного анализа в политологии / Д. В. Нелипа // Вісник Київського національного університету ім. Тараса Шевченка. Філософія. Політологія. — 2012. — № 110. — С. 27–32.
11. Григорьев А. А. Кластерный анализ как инструмент обработки данных при анализе информационных систем / А. А. Григорьев // Известия российского экономического университета им. Г. В. Плеханова. — 2013. — № 1 (11). — С. 036–048.
12. Dap Louw. Forensic Psychology, In International Encyclopedia of the Social & Behavioral Sciences (Second Edition), edited by James D. Wright, Elsevier, Oxford. — 2015. — P. 351–356.
13. Development of the polygraph technique in Japan for detection of deception. Yamamura, Takehiko et al. Forensic Science International. — Vol. 44. — Issue 2. — 257–271.
14. David C. Raskin and John C. Kircher, Chapter 3. — Validity of Polygraph Techniques and Decision Methods, In Credibility Assessment, edited by David C. Raskin and Charles R. Honts John C. Kircher, Academic Press, San Diego. — 2014. — P. 63–129.

УДК 656.212: 004.94

Р. Г. Король,
соиск.;

А. С. Балалаев,
д-р техн. наук, доц.

ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ СИСТЕМЫ «ЖЕЛЕЗНОДОРОЖНАЯ СТАНЦИЯ – МОРСКОЙ ПОРТ» НА ПРИМЕРЕ ВЛАДИВОСТОКСКОГО ТРАНСПОРТНОГО УЗЛА

THE SIMULATION MODELING SYSTEM «RAILWAY STATION – SEAPORT» VLADIVOSTOK TRANSPORT HUB

В данной работе представлена имитационная модель системы «железнодорожная станция – морской порт» для определения максимальной пропускной способности транспортного узла при заданном техническом оснащении обычных условиях и в условиях неравномерности поступления вагонопотока (при