DOI 10.37882/2223–2966.2022.08.18

УСИЛЕНИЕ МЕХАНИЗМОВ ПРОВЕРКИ ГИПОТЕЗЫ НЕЗАВИСИМОСТИ ДАННЫХ

STRENGTHENING MECHANISMS FOR TESTING THE DATA INDEPENDENCE HYPOTHESIS

V. Ziyautdinov
T. Zolotareva
M. Smirnov

In the modern world, it is increasingly necessary to deal with large amounts of information that are included in the concept of Big Data. When processing such information, it is necessary to clearly determine whether there is a relationship between various randomly selected blocks of data. In this paper, we consider various statistical criteria that test the data independence hypothesis and compare them. The application of these criteria allows you to determine whether there is a relationship between different information within big data. In addition, the article tested and analyzed the statistical hypothesis about the independence of data analysis by various numbers of artificial neurons. Subsequently, this will allow creating an optimal self-learning neural network to determine the independence of data when analyzing large amounts of information based on sample data.

Keywords: neuron, statistical criterion, data independence hypothesis.

Зияутдинов Владимир Сергеевич
К.п.н., доцент, Липецкий казачий институт технологии и управления (филиал) ФГБОУ ВО «Московский государственный университет технологии и управления имени К.Г. Разумовского (ПКУ)», г. Липецк
zevslipetsk@yandex.ru
Золотарева Татьяна Александровна
Старший преподаватель, Липецкий государственный педагогический университет им. П.П. Семенова-Тян-Шанского, г. Липецк
zolotarevatatyana2016@yandex.ru
Смирнов Михаил Юрьевич
К.ф.-м.н., доцент, Липецкий казачий институт технологии и управления (филиал) ФГБОУ ВО «Московский государственный университет технологии и управления имени К.Г. Разумовского (ПКУ)», г. Липецк
m_u_smirnov@mail.ru

Аннотация. В современном мире все чаще приходится иметь дело с большими объемами информации, которые входят в понятие Big Data. При обработке такой информации необходимо четко определить: существует ли связь между различными произвольно выбранными блоками данных. В данной работе рассмотрены различные статистические критерии, проверяющие гипотезу независимости данных, и проведено их сравнение. Применение этих критериев позволяет определить имеется ли связь между различной информацией внутри больших данных. Кроме того, в статье проверена и проанализирована статистическая гипотеза о независимости анализа данных различным количеством искусственных нейронов. Впоследствии это позволит создать оптимальную самообучающуюся нейронную сеть для определения независимости данных при анализе больших объемов информации на основании выборочных данных.

Ключевые слова: нейрон, статистический критерий, гипотеза независимости данных.

В процессе статистического анализа иногда бывает необходимо сформулировать и проверить гипотезы относительно рассматриваемых независимых параметров. Так как проверка статистических гипотез осуществляется на основании выборочных данных, т.е. ограниченного ряда наблюдений, решения относительно различных предположений имеют вероятностный характер. Рассмотрим различные статистические критерии проверки независимости данных, проведем сравнительную характеристику.

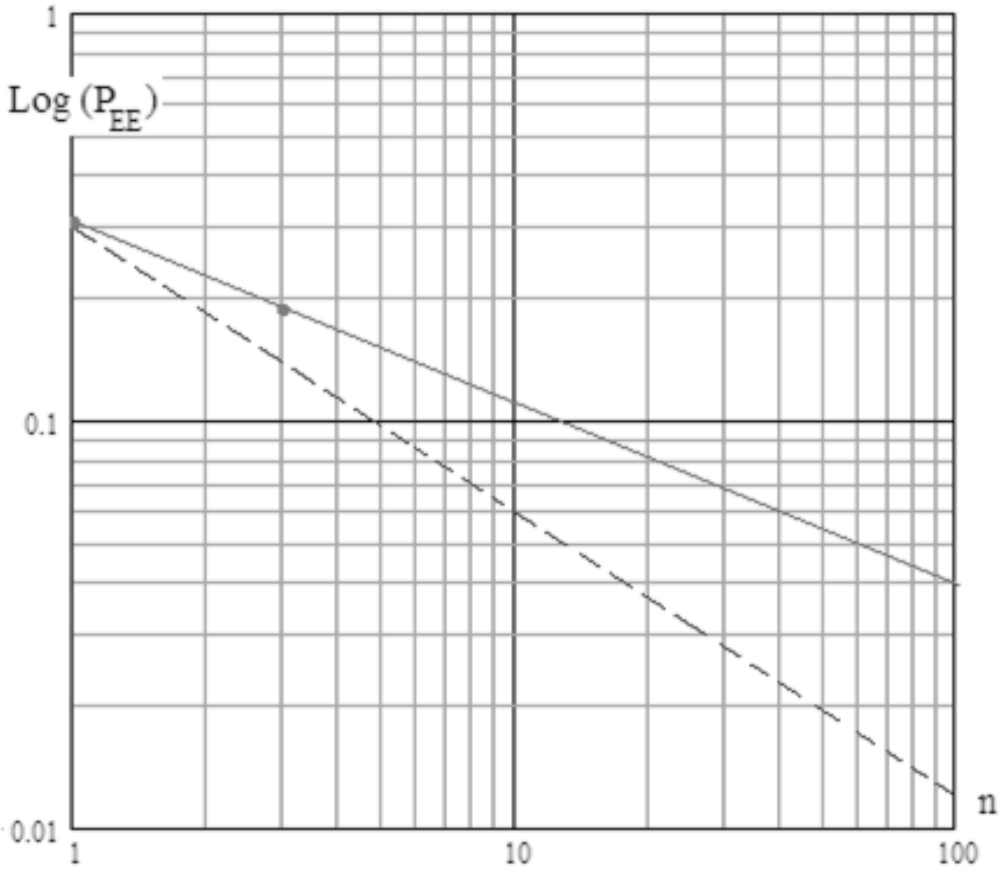

Рис. 1. Предсказание значений равных вероятностей ошибок первого и второго рода для двух настроек механизмов свертывания выходных кодов при разном числе нейронов

Свертывание кода с трехкратной избыточностью, устраняющее неопределенность наименее вероятных кодовых состояний

Проверяя статистическую гипотезу о независимости данных тремя искусственными нейронами, мы фактически решаем некоторую 32-х мерную задачу, рассматривая ее в 32-х мерном пространстве с трех разных ракурсов. При этом кодовое состояние «000» свидетельствует о том, что все три нейрона подтверждают проверяемую гипотезу. Инверсное кодовое состояние «111» свидетельствует об отвержении проверяемой гипотезы всеми нейронами. Воспользуемся простейшим кодом, обнаруживающим и исправляющим ошибочные (наименее вероятные) состояния голосованием по большинству. Этот класс кодов сводится к вычислению расстояний Хэмминга между наблюдаемым кодом и идеальным кодом «000». При расстояниях Хэмминга «h»={«0», «1»}, проверяемая гипотеза независимости принимается. При расстояниях Хэмминга «h»={«2», «3»} гипотеза независимости отвергается.

Численный эксперимент показывает, что расстояния Хэмминга «h=0» появляется с вероятностью 0.519. Выходные коды с расстоянием «h=1» появляются с вероятностью 0.294, то есть рассматриваемый корректирующий код дает подтверждение гипотезы независимости с вероятностью 0.519+0.294=0.813. Тогда значения одинаковых вероятностей ошибок первого и второго рода для молекулы из трех нейронов составит 0.187.

В случае, если в место трех нейронов буден использоваться 21 нейрон, правила принятия решения будут похожими. При расстояниях Хэмминга «h»={«0», «1», …, «10»}, проверяемая гипотеза независимости принимается. При расстояниях Хэмминга h={«11», «12», …, «21»} гипотеза независимости отвергается.

Нейросетевое обобщение трех статистических критериев, легко выполнимо. Воспроизвести соответствующий численный эксперимент несложно для малой выборки в 16 опытов. Естественно, что тот же численный эксперимент для 21 статистического критерия воспроизвести сложнее, однако можно симметризовать

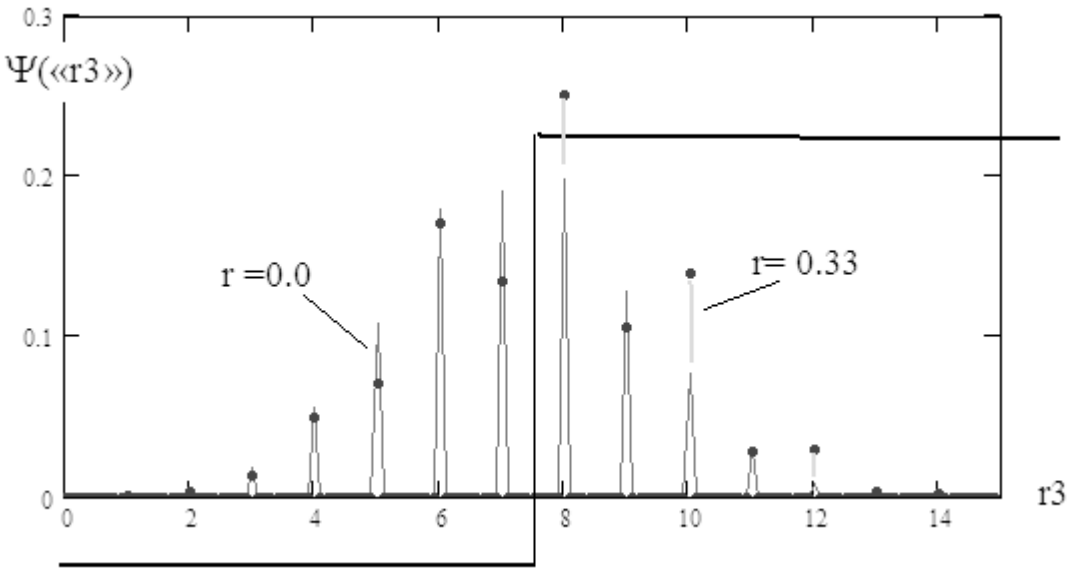

Рис. 2. Дискретный критерий Кенуя и эквивалентный ему нейрон

задачу [1] и получить прогноз ожидаемых вероятностных характеристик для более сложных математических конструкций.

Для выполнения симметризации необходимо вычислить среднее геометрическое вероятностей ошибок первого и второго рода для трех рассматриваемых нейронов:

$$\widetilde{P}_{EE} = \sqrt[3]{0.25 \cdot 0.32 \cdot 0.369} = 0.309 \qquad (1).$$

Далее при симметризации нужно вычислить среднее значение модулей коэффициентов корреляции между откликами накапливающих данные сумматоров трех нейронов:

$$E(|corr(.,.)|) = \frac{|corr(r,r1)| + |corr(r,r2)| + |corr(r1,r2)|}{3} = 0.515 \qquad (2).$$

Если предположить, что вычисленные показатели симметризации по трем нейронам (1) и (2) совпадут с такими же показателями для 21 нейрона, то для предсказания ожидаемых вероятностей ошибок может быть применено линейное приближение, как это отображено на рисунке 1.

Из рисунка 1 видно, что при использовании 21 нейрона вероятность ошибок снижается до величины 0.08. Если удастся довести число параллельно, используемых статистических критериев до 100, то вероятность появления ошибок первого и второго рода $P_1=P_2=P_{EE}$ должна снизиться до величины 0.04. Эта оценка построена на применении простейших кодов корректировки ошибок голосованием по большинству. Если же мы будем применять более сложные алгоритмы обнаружения и исправления ошибок в коде [2], то вероятность ошибок снизится до величины 0.035 уже для 21 нейрона, что вполне приемлемо для практики статистической обработки малых выборок биометрических данных.

Интервал ошибок по сравнению с формулой Пирсона при использовании 21 нейрона сжимается примерно в 7 раз. Это эквивалентно росту объема тестовой выборки в 50 раз или повышению объема исходной выборки с 16 опытов до 800 опытов.

Статистический критерий Кенуя и эквивалентный ему нейрон

Необходимо отметить, что в XX веке были созданы десятки статистических критериев [3], проверяющих гипотезу независимости данных. Например, для этой цели может быть использован критерий Кенуя [4]. Этот критерий построен на анализе рядом стоящих отсчетов выборки с номерами i и (i+1). Если для пара отсчетов {i, i+1} приращения Δx и Δy имеют одинаковый знак, то паре присваивается ранг +1. Расхождение знаков приращений дает отрицательное значение рангов. Сумма рангов по всем парам дает дискретный критерий Кенуя, распределение значений которого иллюстрирует рисунок 2.

Принципиально важным является то, что данные критерия Пирсона и Критерия Кенуя практически ли-

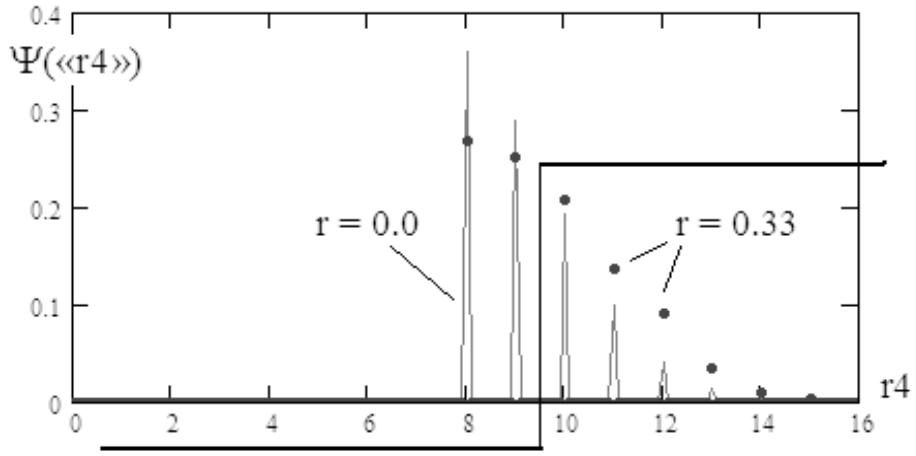

Рис. 3. Критерий Нельсона и эквивалентный ему нейрон

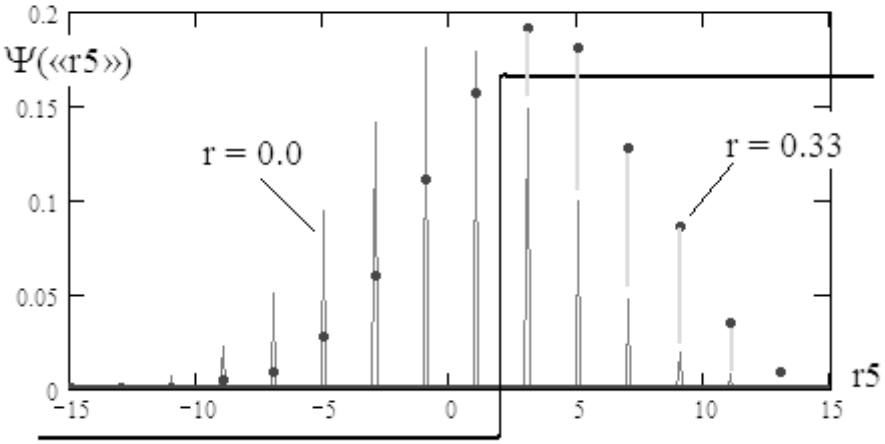

Рис. 4. Модифицированный критерий Нельсона и эквивалентный ему нейрон

нейно независимы corr(r, r3)=0.018. Равновероятная ошибка первого и второго рода составляет P_{EE}=0.442, что вполне пригодно для практического применения этого статистического критерия.

Статистический критерий Нельсона и эквивалентный ему нейрон

Статистический критерий Нельсона создан в 1983 году [5], подобен предыдущему критерию и построен на учете совпадений и не совпадений знаков приращений sign(Δx)=sign(Δy) дает ранг +1, иное соотношение sign(Δx)≠sign(Δy) дает ранг –1. Плюсовые и минусовые ранги подсчитываются. Статистика критерия определяется как минимум или максимум суммы рангов разного знака. Численный эксперимент по моделированию выходных состояний критерия иллюстрируется рисунком 3.

Положительным для этого критерия является его низкий уровень коррелированности с данными корреляционного критерия Пирсона corr(r,«r4»)≈-0.02, однако при этом критерий дает низкий уровень разделимости зависимых и независимых данных по вероятности ошибок первого рода P_1(r=0.0)≈0.351 и по вероятности ошибок второго рода P_2(r=0.33)≈0.519.

Интересно отметить, что использование в качестве статистики разности сумм рангов дает модифицированный критерий Нельсона, спектры состояний которого отображены на рисунке 4.

Для этого варианта критерия Нельсона корреляционная сцепленность с данными нейрона Пирсона увеличивается corr(r,«r5»)≈0.595, а разделимость спектров улучшается P_1(r=0.0)≈0.356, P_2(r=0.33)≈0.362. Это позволяет использовать критерий Нельсона и его модификацию в паре.

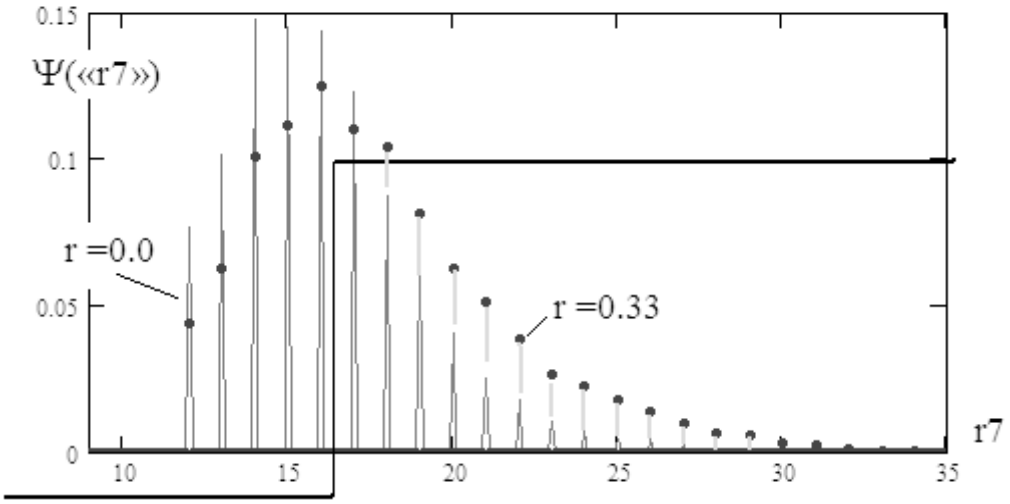

Рис. 5. Статистический нейрон Кокса-Стюарта

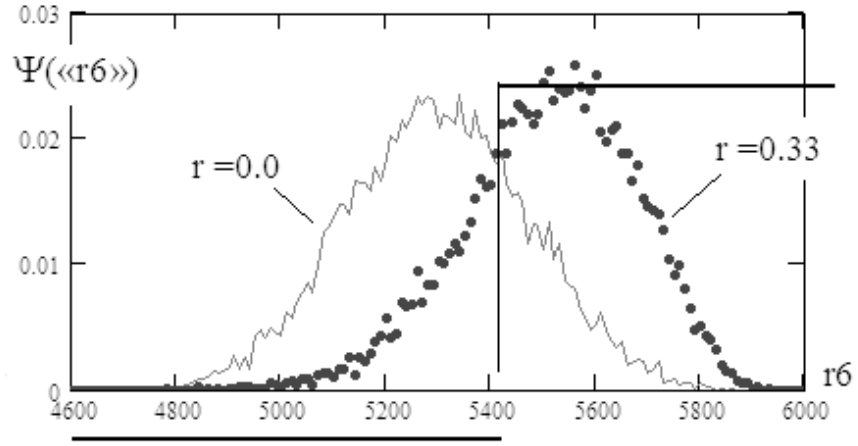

Рис. 6. Критерий Спирмена и эквивалентный ему нейрон

Статистический критерий Кокса-Стюарта и эквивалентный ему нейрон

Статистический критерий Кокса-Стюарта создан 1955 году [6]. Авторами предложено упорядочивать выборку по одной из переменных. Далее авторы предложили разбивать выборку на три или более подвыборок. Для каждой из подвыборок находится минимум и максимум, которые далее используются как пороги для вычисления рангов. В итоговой статистике Кокса-Стюарта все ранги суммируются. Результаты численного эксперимента отражены на рисунке 5.

Из рисунка 5 видно, что значимые амплитуды вероятности спектра критерия Кокса-Стюарта имеют примерно 15 линий. Важным является то, что нейрон Пирсона и нейрон Кокса-Стюарта имеют слабо коррелированные отклики corr(r,«r7»)≈0.004. Кроме того, нейрон Кокса-Стюарта имеет достаточно высокий уровень разделимости $P_1(r=0.0)\approx0.386$, $P_2(r=0.33)\approx0.443$.

Статистический критерий Спирмена и эквивалентный ему нейрон

Статистический критерий Спирмена [7] дискретен, однако он имеет очень большое число спектральных линий даже на малых выборках и по этой причине можно рассматривать его выходные состояния как непрерывные. Состояния нейрона критерия Спирмена отображены на рисунке 6. Состояния плотностей вероятности рисунка 6 практически полностью повторяют

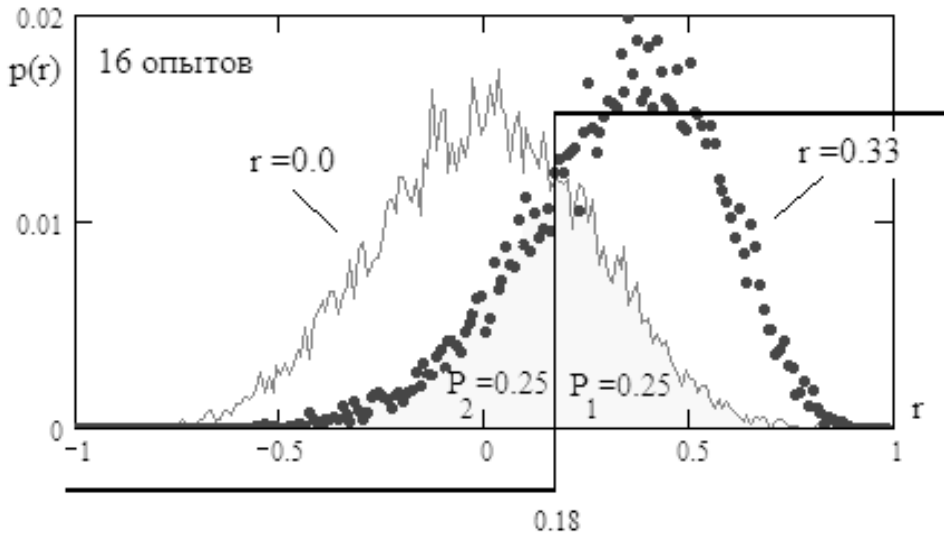

Рис. 7. Вероятности ошибок первого и второго рода корреляционного критерия Пирсона при различении независимых данных малых выборок в 16 опытов и зависимых данных r = 0,33

плотности вероятности рисунка 7 для критерия Пирсона.

Отличие в этих двух распределениях незначительные, различаются только масштабы переменных (непрерывная переменная Пирсона — r изменяется в интервале от –1 до +1, а дискретная переменная Спирмена — «r6» изменяется в интервале от 4800 до 6000).

Корреляционная сцепленность этих двух критериев крайне высока corr (r, «r6»)≈0.928, а показатели разделимости практически тождественны. Все это говорит о том, что нет смысла использовать в паре статистический критерий Пирсона и статистический критерий Спирмена. Эти статистические критерии практически дублируют друг друга.

Нейросетевое обобщение десятков и сотен статистических критериев

Следует отметить, что список перечисленных выше статистических критериев проверки независимости данных может быть существенно расширен. С высокой вероятностью расширенная версия списка критериев может выглядеть следующим образом:

1. Корреляционный критерий Пирсона-Эджуорта-Эудлона (1890 г. [8]);
2. ИНС молекула с квантованием по квандртантам (2017 [9]);
3. ИНС молекула двух эллиптических квантователей (2019 г. [10]);
4. Критерий независимости Кенуя [4];
5. Критерий независимости Нельсона 1983 год [5];
6. Модифицированный критерий Нельсона (Рисунок 4);
7. Критерий Кокса-Стюарта 1955 год [6];
8. Корреляционный критерий Шахани [4];
9. Сериальный критерий Шведа-Эйзерхарта [4];
10. Критерий Спирмена [7];
11. Квандрантный критерий Эландта 1962 год [11];
12. Угловой критерий Олмстеда-Тьюки 1947 год [12]);
13. Критерий Блума-Кифера-Розенблатта 1961 год [13];
14. Ранговый критерий Гётфинга 1948 год [14];
15. Критерий Ширахате 1981 год [15];
16. Критерий корреляции Фишера-Йэйтса 1961 год [16];
17. Критерий конкордации Кенделла-Бебингтона Смита 1939 год [17].

Для первых трех статистических критериев по списку прогноз ожидаемых характеристик для среднего геометрического значения равновероятных ошибок $\widetilde{P}_{EE} \approx 0.309$ и среднее значение модулей коэффициентов корреляции $E(|r|) \approx 0.515$. Результат прогнозирования отражен на рисунке 1. Если учесть влияние друг на друга первых 5 критериев списка, то среднее геометрическое вероятностей ошибок ухудшается $\widetilde{P}_{EE} \approx 0.342$, однако корреляционная сцепленность выходных состояний 5 нейронов падает $E(|r|) \approx 0.311$.

Задавшись конкретным списком, обобщаемых статистических критериев, мы всегда можем вычислить их параметры симметризации, то есть мы заранее можем построить таблицы, предсказывающие ожидаемое качество нейросетевых решений.

ЛИТЕРАТУРА

1. Иванов, А.И. Учет влияния корреляционных связей через их усреднение по модулю при нейросетевом обобщении статистических критериев для малых выборок / А.И. Иванов, А.Г. Банных, Ю.И. Серикова // Надежность».— № 2.— 2020.
2. Безяев, А.В. Биометрико-нейросетевая аутентификация: обнаружение и исправление ошибок в длинных кодах без накладных расходов на избыточность: препринт / А.В. Безяев.— Пенза: Изд-во ПГУ, 2020.— 68 с.
3. Кобзарь, А.И. Прикладная математическая статистика. Для инженеров и научных работников / А.И. Кобзарь.— Москва: ФИЗМАТЛИТ.— 2006.— 816 с.
4. Кенуй, М.Г. Быстрые статистические вычисления. Упрощенные методы оценивания и проверки / М.Г. Кенуй; пер. с англ.— Москва: Статистика, 1979.
5. Nelson, L.S. A sign test for correlation / L.S. Nelson // Journal of Quality Technology, 1983.— Vol. 15, № 4.— P. 199–202.
6. Cox, D.R. Quick tests for trend in location dispersion. / D.R. Cox, A. Stuart // Biometrika, 1955.— Vol. 42.— P. 80–95.
7. Кендэлл, М. Ранговые корреляции / М. Кендэлл; пер. с англ.— Москва: Статистика, 1975.
8. Официальный сайт «Википедия».— URL: https://wikipedia.org/ wiki/Корреляция
9. Волчихин, В.И. Квантовая суперпозиция дискретного спектра состояний математической молекулы корреляции для малых выборок биометрических данных / В.И. Волчихин, А.И. Иванов, А.В. Сериков, Ю.И. Серикова // Вестник Мордовского университета.— Т. 27, № 2.— 2017.— С. 230–243.
10. Сериков, А.В. Корреляционная молекула с эллиптическими квантователями для вычислений на малых обучающих выборках / А.В. Сериков, С.В. Качалин // Безопасность информационных технологий: сб. науч. ст. по материалам I Всерос. науч.-техн. конф., 24 апреля.— Пенза, 2019.— С. 123–129
11. Elandt, R.C. Exact and approximate power of the non-parametric test of tendency / R.C. Elandt // The Annals Mathematical Statistics.— 1962.— Vol. 33.— P. 471–481.
12. Olmstead, P.S. A corner test for association / P.S. Olmstead, J.W. Tukey // The Annals Mathematical Statistics.— 1947.— Vol. 18.— P. 495–513.
13. Blum, J.R. Distribution-free tests of independence based on the sample distribution function / J.R. Blum, J. Kieffer, M. Rosenblatt // The Annals Mathematical Statistics.— 1961.— Vol. 32.— P. 485–498.
14. Hoeffding, W. A non-parametric test of independence / W. Hoeffding // The Annals Mathematical Statistics.— 1948.— Vol. 19.— P. 546–557.
15. Shirahate, S. Infraclass rank tests for independence / S. Shirahate // Biometrika.— 1981.— Vol. 68, № 2.— P. 451–456.
16. Fieller, E.C. Tests for rank correlation coefficients / E.C. Fieller, E.S. Pearson // Biometrika.— 1961.— Vol 48, № 1–2.— P. 29–40.
17. Kendall, M.G. The problem of m rankings / M.G. Kendall, Smith Babington // The Annals Mathematical Statistics.— 1939.— Vol. 10.— P. 275–258.

© Зияутдинов Владимир Сергеевич (zevslipetsk@yandex.ru),
Золотарева Татьяна Александровна (zolotarevatatyana2016@yandex.ru), Смирнов Михаил Юрьевич (m_u_smirnov@mail.ru).
Журнал «Современная наука: актуальные проблемы теории и практики»