

А.Ю. ТИМОФЕЕВА, Т.В. АВДЕЕНКО
**ОТБОР ПРИЗНАКОВ В УСЛОВИЯХ
НЕСБАЛАНСИРОВАННОСТИ КЛАССОВ: ЭМПИРИЧЕСКОЕ
СРАВНЕНИЕ КОРРЕЛЯЦИИ И ВЗАИМНОЙ ИНФОРМАЦИИ**

Тимофеева А.Ю., Авдеенко Т.В. Отбор признаков в условиях несбалансированности классов: эмпирическое сравнение корреляции и взаимной информации.

Аннотация. Рассматривается проблема несбалансированности данных в задачах предварительного отбора признаков с использованием скрининга с гарантированной независимостью. В рамках данного подхода анализируется выбор меры зависимости, применяемой для ранжирования признаков. В работе исследуются два наиболее распространенных показателя взаимосвязи – коэффициент корреляции и взаимная информация. Коэффициент корреляции характеризует линейную зависимость между переменными, его оценки просты в вычислении и обладают хорошими статистическими свойствами. Взаимная информация является более универсальной мерой, способной выявлять нелинейные зависимости, однако ее оценивание требует восстановления распределений по выборочным данным, что приводит к повышенной вычислительной сложности и чувствительности к выбору параметров. На основе серии вычислительных экспериментов показано, что методы оценивания взаимной информации на основе k ближайших соседей уступают ядерным методам с точки зрения устойчивости и качества скрининга. Установлено, что во всех рассмотренных сценариях, включая различные уровни несбалансированности данных, корреляционные меры обеспечивают более стабильные оценки и более надежное ранжирование признаков по сравнению с взаимной информацией. Практическая применимость результатов продемонстрирована на реальных данных электроэнцефалографии, связанных с анализом арифметических способностей. Показано, что двухшаговая процедура, включающая скрининг на основе коэффициента корреляции и последующее моделирование с использованием LASSO-регрессии, обеспечивает существенное снижение ошибки кросс-валидации по сравнению с подходами, основанными на взаимной информации и отсутствием предварительного отбора. Использование главной компоненты, построенной по отобранным признакам, позволяет повысить точность прогноза. Полученные результаты свидетельствуют о том, что в условиях малых выборок, мультиколлинеарности и несбалансированности данных корреляционные меры являются более предпочтительными для задач предварительного отбора признаков с точки зрения устойчивости и воспроизводимости.

Ключевые слова: несбалансированные данные, взаимная информация, коэффициент корреляции, скрининг с гарантированной независимостью, ядерные оценки, k ближайших соседей, LASSO-регрессия.

1. Введение. В задачах высокоразмерного анализа данных, когда число признаков существенно превышает объем выборки, предварительное сокращение размерности является необходимым этапом построения устойчивых статистических моделей. Одним из наиболее распространенных подходов к решению этой задачи является отбор, или скрининг, признаков (feature screening), при котором исходные переменные ранжируются по степени их связи

с откликом, после чего отбирается подмножество наиболее информативных признаков [1].

Применение процедур скрининга позволяет существенно сократить размерность пространства признаков и повысить устойчивость последующего моделирования [2]. Их значимость особенно возрастает в условиях малых выборок [3] и мультиколлинеарности [4], при которых использование полного набора признаков приводит к переобучению и ухудшению обобщающей способности моделей.

Среди методов предварительного отбора признаков широкое распространение получил подход скрининга с гарантированной независимостью (Sure Independence Screening, SIS) [5], в рамках которого каждый признак оценивается независимо от других. Несмотря на концептуальную простоту, данный метод демонстрирует высокую эффективность в задачах с ультравысокой размерностью, особенно в сочетании с процедурами регуляризации, такими как LASSO. Ключевым элементом таких методов является выбор меры зависимости, используемой для ранжирования признаков.

В практике анализа данных в качестве мер зависимости наиболее часто используются коэффициенты корреляции и взаимная информация [6]. Коэффициент корреляции Пирсона применяется для оценки линейной зависимости между количественными переменными. В случае бинарного отклика используется точно-бисериальный коэффициент корреляции [7], представляющий собой естественное обобщение корреляции на случай смешанных типов данных и интерпретируемый как корреляция между непрерывным признаком и индикаторной переменной, кодирующей принадлежность к классу. К достоинствам корреляционных показателей относятся простота вычисления, отсутствие настраиваемых параметров и устойчивость при малых объемах выборки, что делает их особенно удобными для задач скрининга.

Взаимная информация является более общей мерой зависимости, основанной на энтропийных характеристиках распределения. Она позволяет выявлять как линейные, так и нелинейные взаимосвязи без предположений о функциональной форме зависимости. Вместе с тем ее практическое применение связано с необходимостью оценивания по выборочным данным, что может приводить к высокой вариации оценок, особенно при ограниченном объеме выборки.

Несмотря на широкое использование указанных подходов, вопрос о выборе меры зависимости для процедур скрининга остается открытым. Дополнительную сложность представляет

несбалансированность данных, характерная для многих прикладных задач [8]. Она проявляется в существенной асимметрии распределения категориального отклика, при которой миноритарные классы представлены значительно меньшим числом наблюдений. Подобные ситуации возникают, в частности, в задачах обнаружения мошенничества [9], медицинской диагностики [10] и интеллектуального анализа образовательных данных [11]. При этом высокая общая точность классификации не гарантирует адекватного качества предсказания для класса меньшинства [12].

В контексте отбора признаков несбалансированность данных оказывает влияние на точность оценивания мер зависимости, поскольку асимметрия распределения отклика изменяет их статистические свойства. Это может приводить к увеличению дисперсии оценок и снижению надежности ранжирования признаков. Несмотря на значительное количество работ, посвященных обработке несбалансированных данных в задачах классификации [3, 13, 14], влияние данного фактора на выбор меры зависимости для процедур скрининга изучено недостаточно.

В частности, остается открытым вопрос о сравнительной вариативности различных показателей взаимосвязи в условиях асимметричного распределения отклика. Существующие исследования в основном сосредоточены либо на сравнении методов классификации [14], либо на разработке алгоритмов оценивания взаимной информации [15], тогда как систематический анализ пригодности различных мер зависимости именно для задач предварительного отбора признаков представлен ограниченно.

Целью настоящей работы является сравнительный анализ корреляционных и информационно-теоретических мер зависимости в задачах отбора признаков при наличии несбалансированных данных. В работе исследуются статистические свойства оценок, включая их вариацию и устойчивость, а также влияние выбора меры зависимости на качество процедур скрининга.

Для достижения поставленной цели проводится серия модельных экспериментов. В качестве базовой была выбрана линейная модель, позволяющая контролируемо варьировать уровень зависимости и степень несбалансированности данных. Такая постановка позволяет исследовать влияние вариации и смещения оценок независимо от сложности функциональной формы связи. Использование более сложных нелинейных моделей затрудняет интерпретацию результатов, поскольку в этом случае одновременно изменяются как свойства самой зависимости, так и свойства оцениваемых показателей.

Кроме того, выполняется анализ реальных данных электроэнцефалографии (ЭЭГ), характеризующихся высокой размерностью и сложной структурой зависимостей, что позволяет оценить практическую применимость рассматриваемых подходов.

Основной результат работы состоит в том, что в условиях слабых зависимостей, мультиколлинеарности, ограниченного объема выборки и несбалансированности данных корреляционные меры демонстрируют более стабильное и воспроизводимое поведение по сравнению с оценками взаимной информации, несмотря на их меньшую теоретическую универсальность.

Структура работы следующая. В разделе 2 описываются методологические детали рассматриваемых подходов. Раздел 3 посвящен вычислительным экспериментам. В разделе 4 приводятся результаты анализа реальных данных электроэнцефалографии. В заключительных разделах представлены обсуждение полученных результатов и основные выводы.

2. Методы исследования.

2.1. Рассматриваемые меры зависимости и методы их оценивания. В работе рассматриваются два класса показателей взаимосвязи между признаком и откликом: корреляционные и информационно-теоретические меры.

Для количественного отклика используется коэффициент корреляции Пирсона. В случае бинарного отклика применяется точно-бисериальный коэффициент корреляции, который может быть интерпретирован как коэффициент корреляции между непрерывной переменной и бинарным индикатором, кодирующим принадлежность наблюдения к одному из классов.

Следует отметить, что даже в рамках нормальной модели существует множество оценок коэффициента корреляции, среди которых только одна является несмещенной и обладает минимальной дисперсией [16]. Тем не менее на практике, как правило, используется классическая выборочная оценка коэффициента корреляции Пирсона r [16]. Результаты проведенных в работе экспериментов показывают, что ее смещение является незначительным, а дисперсия близка к нижней границе, задаваемой неравенством Рао-Крамера.

Оценка точно-бисериального коэффициента корреляции [2] эквивалентна коэффициенту корреляции Пирсона между непрерывным признаком и бинарной переменной, принимающей значения 0 и 1. Однако данный показатель обладает ограничением: его максимальное абсолютное значение может быть меньше единицы, что зависит

от степени несбалансированности классов [17]. В связи с этим для сопоставления результатов целесообразно использовать бисериальный коэффициент корреляции, характеризующий связь между латентной непрерывной переменной, лежащей в основе бинарного отклика, и наблюдаемым признаком.

Простая оценка r_{pb} бисериального коэффициента может быть получена путем корректировки точно-бисериального коэффициента на множитель, зависящий от степени несбалансированности классов [17]. Альтернативно данный коэффициент может быть оценен методом максимального правдоподобия (обозначим соответствующую оценку как r_b), как описано в [18]. В последнем случае процедура оценивания сводится к решению задачи оптимизации, что повышает вычислительные затраты.

В качестве альтернативы корреляционным мерам рассматривается взаимная информация, позволяющая выявлять как линейные, так и нелинейные зависимости между переменными. В отличие от коэффициентов корреляции, оценивание взаимной информации в непрерывном случае требует задания дополнительных параметров, в частности параметра сглаживания при построении оценок плотностей распределения.

Для оценивания взаимной информации используются два основных подхода [19]. Первый базируется на ядерном оценивании плотностей распределения, при котором взаимная информация вычисляется на основе сглаженных оценок совместного и маргинальных распределений. Данный подход является достаточно гибким и применим как к непрерывным, так и к смешанным данным. Для снижения систематической ошибки может применяться коррекция смещения. В настоящей работе используется реализация данного подхода в пакете *mpmi* среды R [20], где коррекция смещения осуществляется методом складного ножа. Ширина окна для ядерного сглаживания определяется с использованием функции *dpik()* из пакета *KernSmooth*. Обозначим через mi_{ks}^{cc} и mi_{ks}^{cd} соответствующие оценки взаимной информации для непрерывного случая и для случая дискретного отклика.

Второй подход, предложенный в [21], основан на оценивании энтропии с использованием расстояний до k ближайших соседей (k -nearest neighbors). В работе [22] данный метод адаптирован для анализа взаимосвязи между непрерывной и бинарной переменными. Оба варианта реализованы в пакете *knnmi* среды R [23]. Однако исследования свойств оценок для непрерывного отклика,

рассчитанных с помощью функции *mutual_inf_cc()* пакета *knnmi*, показали, что они сильно смещаются к нулю с ростом k , и это смещение не соответствует, описанному в статье [21]. Такие свойства делают их непригодными для скрининга с гарантированной независимостью, поэтому был выбран пакет *YEAB* [24], где функция *mut_info_knn()* дает приемлемое смещение и показывает гораздо лучшие результаты, чем *mutual_inf_cc()* в пакете *knnmi*. Для дискретного отклика использовалась функция *mutual_inf_cd()* пакета *knnmi*.

Существенной особенностью данного подхода является зависимость результатов от выбора параметра k : с увеличением k возрастает систематическая ошибка оценивания, тогда как вариация уменьшается. В [21] рекомендуется выбирать значения параметра k в диапазоне от 2 до 4. Обозначим соответствующие оценки взаимной информации через mi_{knn}^{cc} для непрерывного случая и mi_{knn}^{cd} для случая дискретного отклика.

Рассмотренные подходы широко применяются на практике, однако их чувствительность к выбору параметров и зависимость от структуры данных могут существенно влиять на качество оценивания. Ожидается, что данная чувствительность проявляется особенно существенно в условиях ограниченного объема выборки и несбалансированности классов.

2.2. Процедура скрининга и критерии оценки.

Для предварительного отбора признаков используется процедура скрининга с гарантированной независимостью (Sure Independence Screening, SIS). Данный подход реализует двухшаговую схему. На первом этапе производится фильтрация признаков на основе их индивидуальной взаимосвязи с откликом, что позволяет исключить заведомо нерелевантные переменные. На втором этапе отобранное подмножество анализируется с применением методов регуляризации, в частности LASSO, учитывающих совместное влияние признаков.

Таким образом, ключевую роль на этапе скрининга играют статистические свойства оценок показателей взаимосвязи, используемых для ранжирования признаков, прежде всего их дисперсия и смещение.

Формально [5], в рамках обобщенной линейной модели, связывающей отклик Y с признаками $X_1, \dots, X_p$, процедура SIS предполагает выделение подмножества признаков вида:

$$\hat{M}_{\gamma_n} = \{1 \leq j \leq p: |\hat{\beta}_j^M| \geq \gamma_n\},$$

где $\hat{\beta}_j^M$ – оценка параметра обобщенной линейной модели при переменной X_j , γ_n – заданное пороговое значение, зависящее от объема выборки n . Обычно для удобства на практике задается не γ_n , а число отбираемых переменных $d_n = \left\lfloor \frac{n}{\log(n)} \right\rfloor$. Кроме того, вместо оценок параметров $\hat{\beta}_j^M$ рассматриваются оценки различных показателей взаимосвязи $\hat{\theta}_j$, например, коэффициента корреляции Пирсона, точно-бисериального коэффициента корреляции [2], взаимной информации [6]. Тогда по результатам отбора получаем подмножество:

$$\hat{M}_{d_n} = \{1 \leq j \leq p: \text{первые } d_n \text{ наибольших } |\hat{\theta}_j|\}.$$

В [25] показано, что при надлежащих условиях регулярности эта процедура обладает свойством надежного скрининга:

$$P(M_* \subset M_{\gamma_n}) \rightarrow 1 \text{ при } n \rightarrow \infty,$$

где истинная разреженная модель описывается множеством $M_* = \{1 \leq j \leq p: \beta_j^* \neq 0\}$, β_j^* соответствуют истинным значениям параметров обобщенной линейной модели. Тем самым скрининг асимптотически гарантирует достоверность того, что ни один важный признак не будет потерян при отборе.

В настоящей работе исследуется, в какой степени данное свойство сохраняется в условиях несбалансированности данных и ограниченного объема выборки. Для этого используются два типа критериев. Во-первых, анализируются статистические характеристики оценок показателей взаимосвязи, включая их рассеяние и смещение, при различных уровнях зависимости и степени несбалансированности отклика. Во-вторых, непосредственно оценивается качество процедуры отбора – вероятность включения релевантных признаков в отобранное подмножество.

Дополнительно на реальных данных электроэнцефалографии проводится анализ эффективности двухшаговой процедуры (SIS + LASSO) в задаче прогнозирования арифметических способностей, что позволяет сопоставить теоретические выводы с практическими результатами.

3. Модельные эксперименты.

3.1. План эксперимента. Для анализа статистических свойств рассматриваемых мер зависимости проведена серия модельных экспериментов в контролируемых условиях. Основная цель –

исследовать поведение оценок при варьировании силы связи между переменными и степени несбалансированности классов. Используются упрощенные линейные модели, часто применимые в анализе данных при отсутствии предположений о наличии явной нелинейности и невозможности восстановления зависимостей более сложной формы из-за ограниченности объема выборки.

Рассматривается ситуация, в которой отклик определяется одним релевантным признаком, тогда как остальные признаки являются шумовыми. Генерация данных осуществляется в два этапа.

На первом этапе моделируется непрерывная латентная переменная η , линейно связанная с информативным признаком ξ :

$$\eta = \xi + \varepsilon, \quad (1)$$

где ξ моделировалась как стандартно нормально распределенная величина, ε имеет нормальное распределение с нулевым средним и дисперсией $\rho^{-2} - 1$.

На втором этапе формируется наблюдаемый отклик: либо в виде непрерывной переменной η , либо в виде бинарной переменной ζ , полученной бинаризацией η по порогу, соответствующему выборочному квантилю уровня $P(\zeta = 0)$.

Таким образом, степень несбалансированности классов регулируется порогом бинаризации $P(\zeta = 0)$, что позволяет варьировать распределение отклика от симметричного до существенно асимметричного.

В экспериментах варьируются следующие параметры:

- уровень зависимости $\rho \in [0.01, 0.6]$ с шагом 0.01;
- степень несбалансированности классов $P(\zeta = 0) \in [0.5, 0.95]$

с шагом 0.05.

Для каждого набора параметров проводится многократное моделирование: характеристики оценок усредняются по 10^6 повторениям для корреляционных показателей (в силу их вычислительной простоты) и по 10^4 повторениям для взаимной информации. Для повышения сопоставимости результатов используются одни и те же значения случайных ошибок при различных значениях ρ .

Для моделирования задачи отбора признаков дополнительно рассматривается простая модель:

$$Y = \sum_{j=1}^m f(X_j) + \sum_{k=1}^{p-m} \varepsilon_k, \quad (2)$$

где $X_1, \dots, X_m$ – релевантные признаки с единичной дисперсией, $\varepsilon_1, \dots, \varepsilon_{p-m}$ – нерелевантные признаки с дисперсией 0.001, все признаки распределены нормально и независимо с нулевыми математическими ожиданиями, p – общее число признаков.

Сила связи между откликом и переменными регулировалась путем варьирования значений m, p . Для модели (2) при $f(X_j) = X_j$ справедливо:

$$\rho_{YX_j} = (m + 0.001 \cdot (p - m))^{-1/2}, j = 1, \dots, m, \rho_{Y\varepsilon_k} = \sqrt{0.001} \rho_{YX_1}, \\ k = 1, \dots, p - m,$$

при $f(X_j) = e^{X_j}$ получаем:

$$DY = e(e - 1)m + 0.001 \cdot (p - m), \rho_{YX_j} = \sqrt{e / DY}, j = 1, \dots, m, \\ \rho_{Y\varepsilon_k} = \sqrt{0.001 / DY}, k = 1, \dots, p - m.$$

Рассматриваются три сценария:

1. Умеренная линейная связь: $f(X_j) = X_j$, $m = 5, p = 1000$, $\rho_{YX_j} \approx 0.408, \rho_{Y\varepsilon_k} \approx 0.013$.
2. Слабая линейная связь: $f(X_j) = X_j$, $m = 10, p = 3000$, $\rho_{YX_j} \approx 0.277, \rho_{Y\varepsilon_k} \approx 0.009$.
3. Умеренная нелинейная связь: $f(X_j) = e^{X_j}$, $m = 5, p = 1000$, $\rho_{YX_j} \approx 0.334, \rho_{Y\varepsilon_k} \approx 0.006$.

Объем выборки n фиксируется равным 300. Результаты усредняются по 1000 повторениям. Для моделирования дискретного отклика признак Y в модели (2) был бинаризован с порогом, равным выборочному квантилю уровня $P(\zeta = 0)$.

3.2. Сравнимые показатели взаимосвязи. В рамках экспериментов сравниваются следующие меры зависимости:

- коэффициент корреляции Пирсона r (для непрерывного отклика),
- бисериальный коэффициент корреляции r_b (для бинарного отклика),
- точечно-бисериальный коэффициент корреляции r_{pb} (для бинарного отклика),
- взаимная информация с ядерным оцениванием плотности mi_{ks}^{cc} (для непрерывного отклика) и mi_{ks}^{cd} (для бинарного отклика),

– взаимная информация, оцененная методом ближайших соседей, mi_{knn}^{cc} (для непрерывного отклика) и mi_{knn}^{cd} (для бинарного отклика).

Для методов оценивания взаимной информации параметры сглаживания (ширина окна, число соседей) выбираются из стандартных диапазонов, используемых в прикладных исследованиях, и фиксируются в рамках каждого сценария.

Истинные значения взаимной информации в модели с непрерывным откликом MI^{cont} задавались из известного соотношения [21] для двумерного нормального распределения:

$$MI^{cont}(\rho) = -\frac{1}{2} \log(1 - \rho^2).$$

В случае бинарного отклика значения взаимной информации вычисляются с помощью численного интегрирования по определению [22] с учетом вероятностной структуры модели [18].

Результаты моделирования для случая непрерывного отклика сопоставлялись с теоретической дисперсией эффективных несмещенных оценок коэффициента корреляции для двумерного нормального распределения с неизвестными параметрами сдвига и масштаба [16]:

$$D_{ef}^{cont}(\rho) = \frac{(1-\rho^2)^2}{n}. \quad (3)$$

Это отношение указывает нижнюю границу дисперсии для несмещенных оценок коэффициента корреляции при непрерывном отклике. Сопоставление с ней полученных в ходе экспериментов оценок условного рассеяния показывает, насколько близки оценки r к эффективным.

3.3. Характеристики оценки. Качество рассматриваемых методов оценивается с использованием следующих характеристик:

- оценка условного рассеяния $\widehat{D}_{\hat{\theta}}$ оценки $\hat{\theta}$, характеризующая средний квадрат отклонений оценок от истинного значения параметра θ ;
- оценка условного смещения $\widehat{b}_{\hat{\theta}}$ относительно истинных значений θ , характеризующая систематическую ошибку;
- оценка коэффициента вариации $V\widehat{K}_{\hat{\theta}}$ оценки $\hat{\theta}$ как отношение квадратного корня из рассеяния $\widehat{D}_{\hat{\theta}}$ к истинному значению параметра θ , показывает относительный уровень флуктуации оценок;

– вероятность включения релевантных признаков в отобранное подмножество, характеризующее качество ранжирования признаков.

Особое внимание уделяется области слабых зависимостей, поскольку именно в этом режиме процедуры скрининга наиболее чувствительны к вариативности оценок и риску потери релевантных признаков.

3.4. Результаты. На рисунке 1 представлены результаты моделирования для модели (1) с непрерывным откликом. Для обеспечения сопоставимости методов анализ проводится в терминах коэффициентов вариации, представленных в логарифмической шкале.

Коэффициент корреляции Пирсона демонстрирует практически полное совпадение с теоретической нижней границей вариации, рассчитанной на основе (3), что свидетельствует о его эффективности. Таким образом, выборочная оценка корреляции фактически достигает предельной точности, допустимой в рамках рассматриваемой модели.

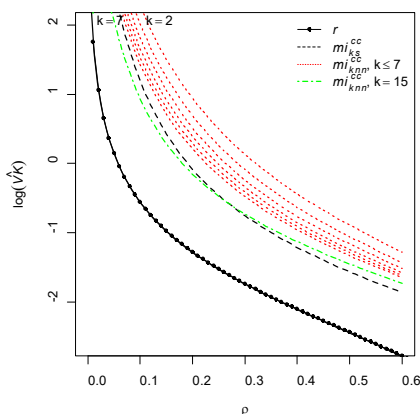


Рис. 1. Оценки коэффициента вариации оценок коэффициента корреляции и взаимной информации, в логарифмической шкале, непрерывный отклик

В отличие от этого, оценки взаимной информации демонстрируют существенно более сложное и менее стабильное поведение, зависящее от используемого метода оценивания. Ядерные оценки характеризуются монотонным снижением вариации с ростом коэффициента корреляции ρ . При этом для малых значений тесноты связи наблюдается значительное положительное смещение, которое уменьшается по мере усиления зависимости и становится отрицательным при больших значениях $\rho \geq 0.55$ (Рис. 2).

Оценки взаимной информации, основанные на расстояниях до k ближайших соседей, демонстрируют меньший коэффициент вариации по сравнению с ядерными оценками вблизи нулевой корреляции ($\rho < 0.25$) при достаточно больших значениях параметра k ($k=15$). Это достигается благодаря тому, что при малых ρ оценки k ближайших соседей имеют меньшую дисперсию и их смещение ближе к нулю (Рис. 2) по сравнению с ядерными оценками.

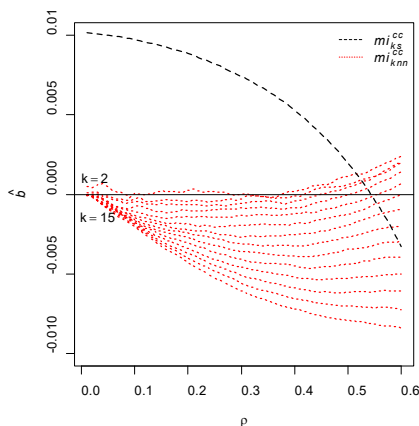


Рис. 2. Оценка смещения оценки взаимной информации, в зависимости от ρ , непрерывный отклик

Для случая бинарного отклика (Рис. 3) наблюдаются аналогичные зависимости. В условиях симметричного распределения отклика увеличение k приводит к снижению рассеяния оценок при слабой зависимости, однако сопровождается ее ростом при более высоких значениях ρ . При существенной несбалансированности данных увеличение k до 20 приводит к снижению вариации на всем диапазоне значений ρ . Тем не менее, во всех рассмотренных сценариях ядерные оценки взаимной информации демонстрируют более устойчивое поведение по сравнению с методами на основе ближайших соседей. Следовательно, при необходимости использования взаимной информации в процедурах скрининга предпочтение следует отдавать ядерным оценкам.

Сравнение оценок бисериального и точноно-бисериального коэффициентов корреляции (Рис. 4) показывает, что их вариативность практически совпадает, особенно в области слабых зависимостей. При этом точноно-бисериальный коэффициент обладает рядом практических преимуществ, включая вычислительную простоту

и более низкую вариативность при малых значениях ρ (вертикальными линиями отмечены значения ρ , ниже которых оценки точно-бисериальной корреляции обладают меньшим рассеянием). Это позволяет рекомендовать его в качестве основного инструмента для скрининга признаков.

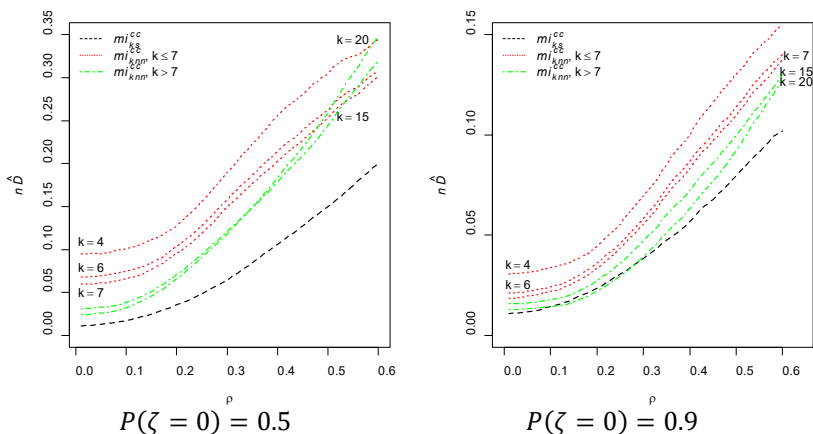


Рис. 3. Оценки рассеяния оценок взаимной информации, умноженные на объем выборки, бинарный отклик

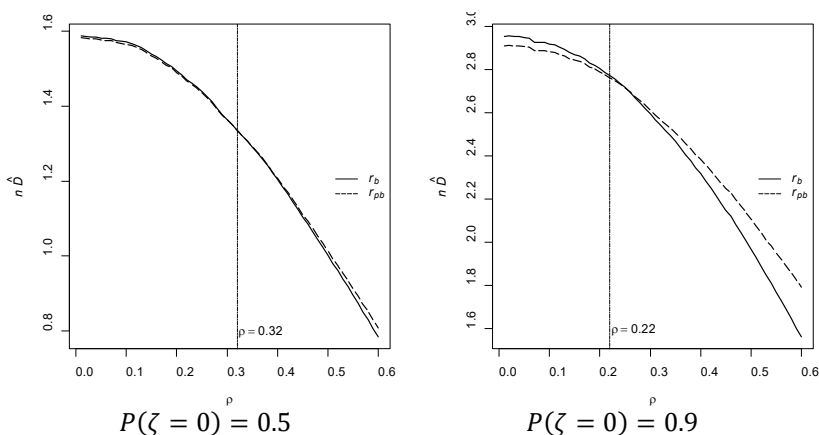


Рис. 4. Оценки рассеяния оценок коэффициента корреляции, умноженные на объем выборки, бинарный отклик

Таким образом, в ходе скрининга с гарантированной независимостью приемлемо использовать точно-бисериальный

коэффициент корреляции и ядерные оценки взаимной информации. Дальнейший анализ направлен на исследование влияния несбалансированности данных. На рисунке 5 представлено отношение коэффициентов вариации для бинарного и непрерывного отклика. Значения ρ взяты с шагом 0.05, стрелками указаны направления уменьшения значений ρ . Установлено, что с ростом асимметрии распределения отклика вариативность оценок возрастает для всех рассматриваемых мер зависимости. Это указывает на принципиальную значимость учета несбалансированности при построении процедур скрининга.

На рисунке 5 прослеживается устойчивая тенденция увеличения отношения коэффициентов вариации (для бинарного отклика по сравнению с непрерывным) по мере роста несбалансированности классов. Это указывает на существенное влияние асимметрии распределения отклика на качество процедур скрининга с гарантированной независимостью.

Для обеих рассматриваемых мер зависимости более слабые связи характеризуются меньшим относительным увеличением вариативности по сравнению с сильными зависимостями. Данный эффект является благоприятным для процедур скрининга, поскольку способствует более надежному исключению нерелевантных признаков.

В сравнении с бисериальной корреляцией ядерные оценки взаимной информации демонстрируют более низкое отношение вариации при переходе от непрерывного отклика к бинарному. Отрицательные значения на рисунке 5 свидетельствуют о том, что в условиях слабой корреляции и умеренной несбалансированности ядерные оценки взаимной информации при бинарном отклике могут обладать относительным преимуществом по сравнению с непрерывным случаем.

Прямое сравнение ядерных оценок взаимной информации mi_{ks}^{cd} и точно-бисериального коэффициента корреляции (Рис. 6) показывает, что взаимная информация систематически уступает корреляции по величине коэффициента вариации (при $\rho \rightarrow 0$ соотношение вариации ядерных оценок взаимной информации и точно-бисериального коэффициента корреляции стремительно растет). Более того, с ростом несбалансированности данных это различие существенно увеличивается. Таким образом, в условиях бинарного отклика использование взаимной информации приводит к менее стабильному ранжированию признаков.

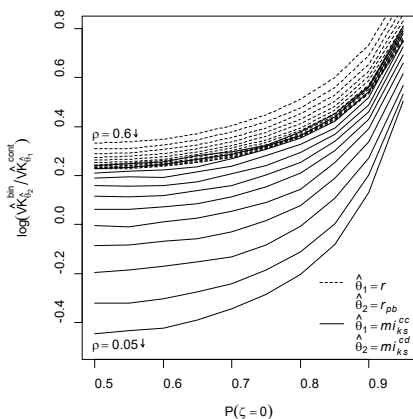


Рис. 5. Отношение оценок коэффициентов вариации ядерных оценок взаимной информации и коэффициентов корреляции при бинарном отклике к случаю непрерывного отклика в зависимости от вероятности нулевого класса, в логарифмической шкале

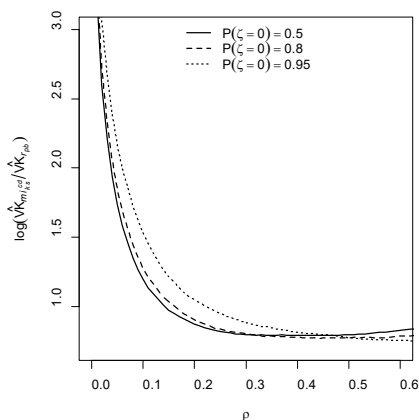


Рис. 6. Отношение оценок коэффициентов вариации оценок MI к ρ , в зависимости от ρ , в логарифмической шкале, бинарный отклик

В процедурах скрининга с гарантированной независимостью ключевым требованием является высокая вероятность включения релевантных признаков в отобранное подмножество (практическая достоверность). Увеличение дисперсии оценок показателей взаимосвязи приводит к снижению этой вероятности, ухудшая качество отбора. Существенную роль при этом играет также смещение оценок. Завышенные оценки увеличивают вероятность включения

нерелевантных признаков, что, в частности, наблюдается для ядерных оценок взаимной информации mi_{ks}^{cc} в области слабых зависимостей ($\rho < 0.55$). В свою очередь, занижение оценок mi_{knn}^{cc} , характерное для методов на основе k ближайших соседей при больших значениях k, приводит к росту вероятности пропуска значимых переменных.

Стоит отметить, что могут быть получены отрицательные оценки взаимной информации. Для целей скрининга они обнулялись.

Анализ качества скрининга подтверждает ранее полученные выводы. Для модели (1) с бинаризацией отклика в качестве порогового значения для коэффициента корреляции фиксируется уровень $\gamma_n = 0.15$, тогда как соответствующие пороги для взаимной информации зависят от доли отрицательного класса и определяются численно на основе ее определения. Предполагается, что истинная корреляция между переменными ξ и η равна 0.3. Вероятность включения релевантного признака $P(|\hat{\theta}| \geq \gamma_n)$ оценивается методом моделирования как доля реализаций, в которых значение выбранного показателя взаимосвязи превышает заданный порог.

Для модели (1) с непрерывным откликом вероятность включения релевантного признака при использовании коэффициента корреляции составляет 0.9969, тогда как для ядерных оценок взаимной информации – 0.992, что соответствует уровню практической достоверности. В то же время методы на основе k ближайших соседей демонстрируют худшие результаты (0.928 при k = 20).

Результаты для бинарного отклика (таблица 1) показывают, что с увеличением несбалансированности данных вероятность включения релевантного признака заметно снижается и перестает соответствовать уровню практической достоверности. Бисериальный и точечно-бисериальный коэффициенты корреляции дают практически совпадающие результаты, что позволяет отдать предпочтение точечно-бисериальному коэффициенту ввиду его вычислительной простоты.

Ядерные оценки взаимной информации существенно превосходят методы на основе расстояний до k ближайших соседей, однако уступают корреляционным мерам. В частности, при высокой степени несбалансированности (вероятность положительного класса 0.1) вероятность пропуска релевантного признака при использовании ядерных оценок взаимной информации составляет 0.122, тогда как для бисериальной корреляции она более чем в два раза ниже и равна 0.056.

На первом этапе скрининга с гарантированной независимостью для модели (2) пороговые значения показателей взаимосвязи γ_n задавались исходя из фиксированного размера

отбираемого подмножества признаков $d_n = \left\lfloor \frac{n}{\log(n)} \right\rfloor = 52$. При слабой линейной зависимости (сценарий В) и использовании коэффициента корреляции Пирсона все релевантные признаки включались в отобранное подмножество в 92.6% случаев, тогда как при умеренной линейной зависимости (сценарий А) – во всех реализациях, при нелинейной зависимости (сценарий С) – в 99.4% случаев. Это указывает на то, что при преобладании слабых связей в исходном наборе признаков даже скрининг на основе коэффициента корреляции не обеспечивает гарантированного включения всех значимых переменных.

Таблица 1. Оценка $P(|\hat{\theta}| \geq \gamma_n)$ при различных методах оценивания показателей взаимосвязи и разном уровне несбалансированности данных

Метод оценивания	$P(\zeta = 0)$				
	0.5	0.6	0.7	0.8	0.9
r_b	0.985	0.983	0.980	0.971	0.944
r_{pb}	0.985	0.983	0.980	0.971	0.943
mi_{ks}^{cd}	0.958	0.956	0.950	0.931	0.878
$mi_{knn}^{cd}, k=4$	0.795	0.798	0.793	0.786	0.756
$mi_{knn}^{cd}, k=6$	0.834	0.833	0.838	0.833	0.790
$mi_{knn}^{cd}, k=15$	0.913	0.911	0.914	0.890	0.798
$mi_{knn}^{cd}, k=20$	0.929	0.927	0.923	0.897	0.786

Для сценария В увеличение размера отбираемого подмножества приводит к заметному улучшению результатов. В частности, при увеличении размера до $d_n = 2 \left\lfloor \frac{n}{\log(n)} \right\rfloor = 104$ и до $d_n = 3 \left\lfloor \frac{n}{\log(n)} \right\rfloor = 156$ (в соответствии с подходом [2]) вероятность включения всех релевантных признаков возрастает до 0.971 и 0.985 соответственно, что может рассматриваться как практически приемлемый уровень.

Для сценария С при увеличении размера отбираемого подмножества вероятность включения всех релевантных признаков на основе коэффициента корреляции повышается: при $d_n = 104$ до 0.997, при $d_n = 156$ до 0.998. Для такого сценария с нелинейной зависимостью ядерные оценки взаимной информации проявляют себя еще лучше: при $d_n = 52$ вероятность включения всех релевантных признаков оценена как 0.999, а при $d_n = 104$ уже составляет 1.

В остальных сценариях использование ядерных оценок взаимной информации показывает существенно менее устойчивые результаты. В случае умеренной линейной зависимости доля реализаций, в которых все релевантные признаки включены в подмножество из 52 переменных,

составляет 0.991, тогда как при слабой зависимости она резко снижается до 0.013. Увеличение размера подмножества до 156 приводит лишь к незначительному росту этого показателя (до 0.107) в условиях слабой связи. Таким образом, в сценарии В скрининг с гарантированной независимостью на основе ядерных оценок взаимной информации оказывается практически непригодным.

Результаты для модели (2) при бинаризации отклика представлены в таблице 2, где $P\left(\bigcap_{j=1}^m \left|\hat{\theta}_{YX_j}\right| \geq \gamma_n\right)$ обозначает вероятность того, что оценки показателей взаимосвязи одновременно для всех релевантных признаков превышают заданный порог. Для моделей А и С в условиях умеренной зависимости получены приемлемые результаты. Точечно-бисериальный коэффициент корреляции обеспечивает практически достоверное включение всех релевантных признаков даже при очень высокой степени асимметрии распределения отклика. При низкой доле положительного класса (0.1) ядерные оценки взаимной информации демонстрируют неудовлетворительные результаты в модели А, тогда как в условиях нелинейных зависимостей (модель С) плохой результат дает и точечно-бисериальный коэффициент корреляции. Как и следовало ожидать, увеличение степени несбалансированности приводит к ухудшению качества скрининга.

Несмотря на умеренные нелинейные зависимости в модели С при достаточно сбалансированных данных (вероятность отрицательного класса от 0.5 до 0.6) точечно-бисериальный коэффициент корреляции обеспечивает такое же качество скрининга, как и взаимная информация. При увеличении вероятности отрицательного класса до 0.8 наблюдаются незначительные преимущества взаимной информации, тогда как при 0.9 – различия в пользу точечно-бисериального коэффициента корреляции, но они очень малы. Следовательно, при умеренной нелинейной связи скрининг на основе коэффициента корреляции предпочтителен, поскольку обеспечивает сопоставимое с взаимной информацией качество отбора, но вычислительно гораздо менее затратный.

В целом результаты моделирования выявляют устойчивые различия в поведении рассматриваемых мер зависимости. Корреляционные показатели обеспечивают низкую вариативность оценок и стабильное ранжирование признаков в условиях умеренной зависимости, включая сценарии с высокой несбалансированностью данных. В то же время при слабых зависимостях эффективность скрининга снижается даже для корреляционных мер.

Напротив, методы оценивания взаимной информации характеризуются существенно большей вариативностью, особенно в области слабых зависимостей, причем данный эффект усиливается с ростом несбалансированности данных.

Таблица 2. Оценка $P\left(\bigcap_{j=1}^m \left|\hat{\theta}_{YX_j}\right| \geq \gamma_n\right)$ при различных методах оценивания показателей взаимосвязи и разном уровне несбалансированности данных

Метод оценивания	Сценарий, d_n	$P(\zeta = 0)$				
		0.5	0.6	0.7	0.8	0.9
r_{pb}	A, 52	1	0.999	0.998	0.996	0.95
r_{pb}	A, 104	1	1	1	1	0.984
r_{pb}	A, 156	1	1	1	1	0.994
mi_{ks}^{cd}	A, 52	0.982	0.983	0.974	0.92	0.603
mi_{ks}^{cd}	A, 104	0.995	0.993	0.991	0.963	0.787
mi_{ks}^{cd}	A, 156	0.998	0.998	0.996	0.981	0.865
r_{pb}	B, 52	0.406	0.375	0.27	0.121	0.004
r_{pb}	B, 104	0.642	0.603	0.509	0.273	0.025
r_{pb}	B, 156	0.743	0.709	0.627	0.406	0.081
mi_{ks}^{cd}	B, 52	0.015	0.016	0.002	0	0
mi_{ks}^{cd}	B, 104	0.073	0.067	0.024	0.009	0
mi_{ks}^{cd}	B, 156	0.139	0.115	0.06	0.022	0
r_{pb}	C, 52	1	0.998	0.993	0.951	0.652
r_{pb}	C, 104	1	1	0.997	0.974	0.795
r_{pb}	C, 156	1	1	1	0.984	0.856
mi_{ks}^{cd}	C, 52	1	0.998	0.995	0.962	0.655
mi_{ks}^{cd}	C, 104	1	1	0.998	0.976	0.781
mi_{ks}^{cd}	C, 156	1	1	0.999	0.988	0.847

В целом проведенные эксперименты выявляют устойчивое преимущество корреляционных мер зависимости. Они обеспечивают низкую вариативность оценок и высокую вероятность корректного отбора релевантных признаков в широком диапазоне условий. Взаимная информация, несмотря на теоретическую универсальность, характеризуется повышенной вариативностью и чувствительностью к параметрам оценивания, что приводит к снижению качества скрининга, особенно в условиях слабых зависимостей и несбалансированных данных.

3.5. Выводы. Проведенные вычислительные эксперименты позволяют сделать следующие обобщающие выводы.

Во-первых, корреляционные меры зависимости демонстрируют наименьшую вариативность оценок и обеспечивают наиболее стабильное ранжирование признаков в широком диапазоне параметров, включая условия слабых зависимостей.

Во-вторых, оценки взаимной информации характеризуются существенно большей вариативностью, особенно в области слабых зависимостей, что приводит к снижению вероятности включения релевантных признаков в процедуры скрининга.

В-третьих, методы оценивания взаимной информации на основе k ближайших соседей проявляют высокую чувствительность к выбору параметра k , вследствие чего уступают ядерным методам и не обеспечивают надежного отбора признаков.

В-четвертых, несбалансированность данных приводит к росту вариативности оценок для всех рассматриваемых мер зависимости, однако данный эффект значительно сильнее выражен для взаимной информации, что дополнительно снижает эффективность соответствующих процедур скрининга.

4. Исследование на реальных данных: анализ ЭЭГ

4.1. Описание данных. Для демонстрации практической применимости рассмотренных подходов используется набор данных ЭЭГ, характеризующийся высокой размерностью признакового пространства и сложной структурой зависимостей между признаками.

Данные заимствованы из [26] и содержат записи ЭЭГ, а также оценки арифметических способностей 36 испытуемых, полученные в процессе выполнения когнитивных заданий. Таким образом, анализ проводится в условиях малого объема выборки.

Предварительная обработка данных выполнена в работе [27]. В качестве признаков использованы метрики связности графов, построенных на основе пяти методов оценки функциональной коннективности (Coh – когерентность, ImCoh – мнимая когерентность, PLI – индекс фазового лага, DPLI – направленный индекс фазового лага, WPLI – взвешенный индекс фазового лага) в пяти частотных диапазонах $(\delta, \theta, \alpha, \beta, \gamma)$. Для каждой комбинации метода и диапазона формировался граф коннективности, определяемый значимыми взаимодействиями между областями мозга, после чего для каждого графа вычислялись семь метрик связности: APL, CPL – средняя и характерная длина пути, D – диаметр, CC – коэффициент кластеризации, Mod – модулярность, EC – центральность по собственному вектору, Cl – центральность по близости. Таким образом, исходно формировалось $5 \times 5 \times 7 = 175$ признаков.

Дополнительно признаки рассчитывались для девяти уровней порогов значимости (квантилей от 0.1 до 0.9 абсолютных значений показателей коннективности для каждого индивида), что увеличило размерность признакового пространства до 1575 переменных. При этом для малых значений порога отдельные метрики связности (в частности, средняя и медианная длина пути, диаметр) принимали практически постоянные значения. Для предотвращения искажения оценок взаимосвязи из анализа были исключены признаки, в которых одно значение встречалось с частотой 60% и более. В результате итоговая размерность пространства признаков составила 1141.

Таким образом, число признаков существенно превышает объем выборки, что делает задачу особенно чувствительной к качеству предварительного отбора.

Преимуществом данного набора данных является возможность рассмотрения двух типов отклика. В качестве непрерывного отклика используется показатель арифметической продуктивности – число правильно решенных заданий за единицу времени. Кроме того, на основе этого показателя формируется бинарный отклик, отражающий уровень математических способностей: испытуемые классифицируются в зависимости от того, превышает ли число решенных задач пороговое значение 10. Доля испытуемых с результатом, не превышающим данный порог, составляет 27.7%, что соответствует выраженной несбалансированности классов.

Перед проведением анализа данные были стандартизованы.

Используемые признаки представляют собой агрегированные характеристики графов функциональной коннективности мозга. Для них линейные аппроксимации широко применяются в исследованиях индивидуальных различий и когнитивных способностей, что подтверждается работами, посвященными анализу связности ЭЭГ [28, 29]. Поэтому применение известной двухшаговой процедуры скрининга с гарантированной независимостью [2] на основе линейной модели с регуляризацией представляется оправданным.

4.2. План исследования. На первом этапе выполняется скрининг признаков с использованием различных мер зависимости: точно-бисериального коэффициента корреляции и взаимной информации, оцененной ядерным методом. Признаки ранжируются по величине соответствующего показателя, после чего отбирается подмножество наиболее значимых переменных. Пороговые значения параметров взаимосвязи γ_n задаются косвенно – через фиксированный размер отбираемого подмножества признаков $d_n = \left\lfloor \frac{n}{\log(n)} \right\rfloor = 10$.

На втором этапе на отобранных признаках строится предсказательная модель с использованием LASSO-регрессии, позволяющей учитывать совместное влияние переменных и выполнять регуляризацию. Поскольку исходное пространство признаков характеризуется выраженной мультиколлинеарностью [27], предварительно на основе отобранных признаков строятся главные компоненты. Это обеспечивает ортогональность входных факторов и повышает устойчивость оценивания, одновременно позволяя учитывать структуру взаимосвязей между признаками. Использование главных компонент также приводит к дополнительному снижению размерности, что важно для интерпретируемости результатов.

4.3. Метрики оценивания. Оценка качества моделей проводится с использованием процедуры перекрестной проверки по принципу «leave-one-out» (LOOCV), что позволяет минимизировать риск переобучения при малом объеме выборки.

Для непрерывного отклика в рамках LOOCV рассчитывается среднеквадратическая ошибка (MSE). Для бинарного отклика используется обобщенная линейная модель с LASSO-регуляризацией, и в качестве метрики качества рассматривается среднее отклонение, вычисляемое по результатам LOOCV:

$$Deviance = -\frac{2}{n} \sum_{i=1}^n y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i),$$

где y_i – наблюдаемое значение бинарного отклика, а $\hat{p}_i$ – оценка вероятности принадлежности i -го объекта к положительному классу, полученная по модели, обученной без учета данного наблюдения.

Дополнительно анализируется устойчивость процедуры отбора признаков при варьировании размера отобранного подмножества, а также при использовании различных мер зависимости.

4.4. Результаты. На рисунке 7 представлены зависимости среднеквадратической ошибки от параметра регуляризации λ для моделей без предварительного отбора признаков (all), а также с отбором на основе коэффициента корреляции Пирсона и ядерных оценок взаимной информации. Штриховыми линиями соответствующего цвета отображены отклонения от среднего на одну стандартную ошибку.

Во всех случаях предварительный скрининг приводит к снижению ошибки по сравнению с использованием полного набора признаков. Однако значимое улучшение по сравнению с базовой моделью (без признаков) достигается только при использовании

корреляционного критерия: в этом случае ошибка снижается почти в два раза. При выборе параметра регуляризации по правилу одного стандартного отклонения итоговая модель включает лишь одну главную компоненту, которая демонстрирует сильную положительную связь с откликом (коэффициент корреляции Пирсона равен 0.781).

Стабильность результатов скрининга на основе коэффициента корреляции исследована при варьировании размера подмножества признаков d_n в диапазоне от 10 до 2 (Рис. 8). Для интерпретации вклада отдельных признаков ($\hat{\beta}_j^M$) оценки коэффициентов LASSO-регрессии при главной компоненте пересчитывались с учетом нагрузок исходных признаков. Полученные результаты свидетельствуют о том, что ряд характеристик структуры графа коннективности, в частности коэффициент кластеризации CC [27], положительно связан с арифметическими способностями.

Для бинарного отклика (Рис. 9) получены качественно сходные результаты при варьировании параметра регуляризации λ . Предварительный скрининг также снижает значение функции потерь *Deviance* по сравнению с использованием всех признаков (all). Однако для ядерных оценок взаимной информации улучшение достигается только при включении нескольких главных компонент (3-4), что существенно усложняет интерпретацию модели. В то же время при использовании точно-бисериального коэффициента корреляции минимум отклонения достигается уже при одной компоненте.

Анализ устойчивости оценок влияния признаков в бинарном случае (рисунок 10) показывает, что знаки эффектов распределяются примерно поровну между положительными и отрицательными. При этом наблюдается согласованность с непрерывным случаем для ряда характеристик: например, диаметр графа на основе мнимой когерентности в α -диапазоне при пороге значимости 0.5 ($D \propto ImCoh 0.5$) и модулярность ($Mod \propto DPLI 0.7$) оказывают отрицательное влияние, тогда как центральность по собственному вектору ($EC \propto DPLI$) на основе направленного индекса фазового лага в δ -диапазоне – положительное.

В целом результаты показывают, что выбор меры зависимости оказывает существенное влияние на итоговое качество модели. Скрининг на основе корреляции обеспечивает более низкую прогнозную ошибку, а также демонстрирует большую устойчивость к вариациям выборки. В то же время использование взаимной информации, хотя и улучшает результаты по сравнению с отсутствием отбора, не приводит к значимому снижению ошибки относительно базового уровня.

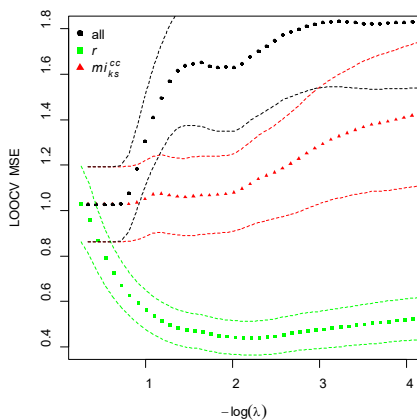


Рис. 7. Среднеквадратическая ошибка LASSO-регрессии при разных вариантах отбора признаков

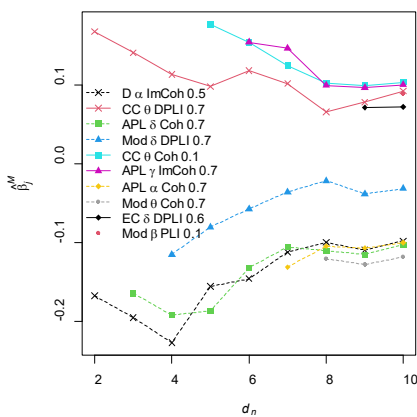


Рис. 8. Результаты скрининга с гарантированной независимостью на основе коэффициента корреляции Пирсона

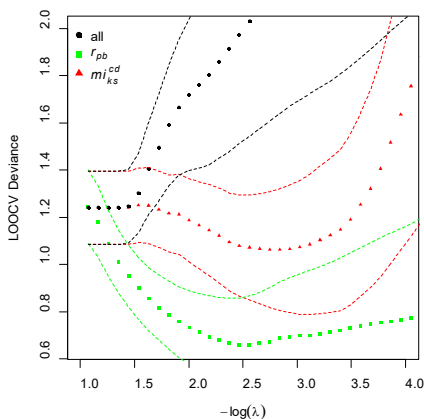


Рис. 9. Среднее отклонение обобщенной линейной LASSO-регрессии при разных вариантах отбора признаков

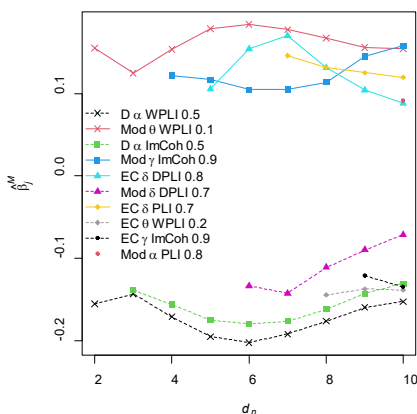


Рис. 10. Результаты скрининга с гарантированной независимостью на основе точно-бисериального коэффициента корреляции

4.5. Выводы. Полученные результаты подтверждают выводы модельных экспериментов и позволяют сформулировать следующие практические положения.

Во-первых, в условиях высокой размерности признакового пространства и несбалансированности данных корреляционные меры являются более надежным инструментом для предварительного отбора признаков, обеспечивая устойчивые и воспроизводимые результаты.

Во-вторых, применение взаимной информации требует осторожности, поскольку процедуры ее оценивания являются

вычислительно более сложными и чувствительными к выбору параметров, что снижает устойчивость результатов.

В-третьих, преимущества взаимной информации как универсальной меры зависимости не всегда реализуются на практике, особенно при малом объеме выборки, слабых зависимостях и выраженной несбалансированности данных.

5. Обсуждение. Полученные результаты демонстрируют согласованную картину как на модельных, так и на реальных данных: выбор меры зависимости оказывает существенное влияние на эффективность процедур скрининга признаков, особенно в условиях слабых зависимостей и несбалансированности данных.

Ключевым результатом является устойчиво более высокая вариативность оценок взаимной информации по сравнению с корреляционными показателями. Это приводит к снижению надежности ранжирования признаков и увеличивает вероятность пропуска релевантных переменных при использовании скрининга с гарантированной независимостью. Данный эффект особенно критичен в задачах, где решения принимаются на основе малых различий между слабыми сигналами.

Наблюдаемые различия объясняются особенностями процедур оценивания. Взаимная информация требует дополнительного этапа оценивания распределений, что связано с выбором настраиваемых параметров (например, числа ближайших соседей или ширины окна в ядерных методах). Эти параметры определяют компромисс между смещением и дисперсией и делают оценки чувствительными к объему выборки и структуре данных. В отличие от этого, корреляционные меры имеют аналитические оценки и не требуют настройки, что обеспечивает их более стабильное поведение.

Несбалансированность данных усиливает выявленные различия. Асимметрия распределения отклика приводит к увеличению неопределенности оценок взаимной информации, тогда как корреляционные показатели оказываются менее чувствительными к данному фактору. Это согласуется с результатами моделирования, где рост степени несбалансированности сопровождался увеличением вариативности оценок взаимной информации и снижением качества скрининга.

Полученные результаты показывают, что теоретическая универсальность взаимной информации не всегда трансформируется в практические преимущества в задачах предварительного отбора признаков. В условиях ограниченного объема выборки и слабых зависимостей более простые меры могут обеспечивать более надежные

и воспроизводимые результаты. Это подчеркивает необходимость учета статистических свойств оценок при выборе методов анализа.

В то же время следует отметить ограничения проведенного исследования. Рассмотренные модели охватывают широкий, но не исчерпывающий класс зависимостей: линейные либо адекватно аппроксимируемые линейной моделью. В частности, при наличии выраженных нелинейных эффектов взаимная информация потенциально может обладать преимуществами. Кроме того, результаты могут зависеть от конкретных методов оценивания и выбора параметров. Эти ограничения требуют проведения дополнительных исследований, выходящих за рамки настоящей работы.

6. Заключение. В работе проведен сравнительный анализ корреляционных и информационно-теоретических мер зависимости в задачах скрининга признаков при наличии несбалансированных данных.

На основе результатов модельных экспериментов и анализа реальных данных показано, что корреляционные меры обеспечивают более стабильные и воспроизводимые оценки в условиях слабых зависимостей, ограниченного объема выборки и асимметричного распределения отклика. Взаимная информация, несмотря на свою теоретическую универсальность, характеризуется повышенной вариативностью и чувствительностью к параметрам оценивания, что приводит к снижению эффективности процедур предварительного отбора признаков.

Полученные результаты указывают на то, что при выборе меры зависимости для задач скрининга необходимо учитывать не только способность выявлять сложные зависимости, но и статистическую устойчивость оценок. В рассматриваемых условиях простые корреляционные показатели оказываются предпочтительными с точки зрения надежности и воспроизводимости результатов.

Дальнейшие исследования могут быть направлены на анализ более сложных типов зависимостей и на разработку методов оценивания взаимной информации, обладающих повышенной устойчивостью в условиях малых выборок и несбалансированных данных.

Литература

1. Liu J.Y., Zhong W., Li R.Z. A selective overview of feature screening for ultrahigh-dimensional data // Science China Mathematics. 2015. vol. 58. pp. 1–22. DOI: 10.1007/s11425-015-5062-9.
2. Jiang J., Shang J. Feature Screening for High-Dimensional Variable Selection in Generalized Linear Models // Entropy. 2023. vol. 25. no. 6. 851 p. DOI: 10.3390/e25060851.

3. Wasikowski M., Chen X. Combating the Small Sample Class Imbalance Problem Using Feature Selection // *IEEE Transactions on Knowledge and Data Engineering*. 2010. vol. 22. no. 10. pp. 1388–1400. DOI: 10.1109/TKDE.2009.187.
4. Chan J.Y.-L., Leow S.M.H., Bea K.T., et al. Mitigating the Multicollinearity Problem and its Machine Learning Approach: A Review // *Mathematics*. 2022. vol. 10. no. 8. 1283 p. DOI: 10.3390/math10081283.
5. Fan J., Fan Y. High Dimensional Classification Using Features Annealed Independence Rules // *Annals of Statistics*. 2008. vol. 36. no. 6. pp. 2605–2637. DOI: 10.1214/07-AOS504.
6. Xu Y., Qian Q. i-SISSO: Mutual information-based improved sure independent screening and sparsifying operator algorithm // *Engineering Applications of Artificial Intelligence*. 2022. vol. 116. 105442 p. DOI: 10.1016/j.engappai.2022.105442.
7. Shukla G., Singh S., Dhule C., et al. Point biserial correlation symbiotic organism search nanoengineering based drug delivery for tumor diagnosis // *Scientific Reports*. 2024. vol. 14. 6530 p. DOI: 10.1038/s41598-024-55159-6.
8. Huyen C. Designing Machine Learning Systems // Sebastopol: O'ReillyMedia. 2022. 443 p.
9. Mqadi N., Naicker N., Adeliyi T. Solving Misclassification of the Credit Card Imbalance Problem Using Near Miss // *Mathematical Problems in Engineering*. 2021. no. 1. 1–16 pp. DOI: 10.1155/2021/7194728.
10. Tasci E., Zhuge Y., Camphausen K., Krauze, A.V. Bias and Class Imbalance in Oncologic Data – Towards Inclusive and Transferrable AI in Large Scale Oncology Data Sets // *Cancers*. 2022. vol. 14. no. 12. 2897 p. DOI: 10.3390/cancers14122897.
11. Wongvorachan T., He S., Bulut O. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining // *Information*. 2023. vol. 14. no. 1. 54 p. DOI: 10.3390/info14010054.
12. Kumar A., Goel S., Sinha N., Bhardwaj A. A Review on Unbalanced Data Classification // *Proceedings of International Joint Conference on Advances in Computational Intelligence (IJCACI 2021)*. 2022. pp. 197–208. DOI: 10.1007/978-981-19-0332-8_14.
13. Mladenic D., Grobelnik M. Feature Selection for Unbalanced Class Distribution and Naive Bayes // *Proceedings of the Sixteenth International Conference on Machine Learning (ICML '99)*. 1999. pp. 258–267.
14. Wang L., Han M., Li X., et al. Review of Classification Methods on Unbalanced Data Sets // *IEEE Access*. 2021. vol. 9. pp. 64606–64628. DOI: 10.1109/ACCESS.2021.3074243.
15. Belghazi M.I., Baratin A., Rajeshwar S., et al. Mutual Information Neural Estimation // *Proceedings of the 35th International Conference on Machine Learning (PMLR)*. 2018. vol. 80. pp. 531–540.
16. Джонсон Н.Л., Коц С., Балакришнан Н. Одномерные непрерывные распределения: ч. 2 / Пер. 2-го англ. изд. Волковой О.И., Стригуновой М.С., Шиховой Н.А. / Под ред. Чепурина Е.В. // М.: БИНОМ. Лаборатория знаний. 2012. 600 с.
17. Cheng Y., Liu H. A short note on the maximal point-biserial correlation under non-normality // *British Journal of Mathematical and Statistical Psychology*. 2016. vol. 69. no. 3. pp. 344–351. DOI: 10.1111/bmsp.12075.
18. Timofeeva A.Y. Data Type Detection for Choosing an Appropriate Correlation Coefficient in the Bivariate Case // *Proceedings of the Second Young Scientist's International Workshop on Trends in Information Processing (YSIS2)*. 2017. vol. 1837. pp. 188–194.
19. Czyz P., Grabowski F., Vogt J., Beerenwinkel N., Marx A. Beyond Normal: On the Evaluation of Mutual Information Estimators // *Advances in Neural Information Processing Systems*. 2023. vol. 36. pp. 1–34. DOI: 10.52202/075280-0742.

20. Mpmi: Mixed-Pair Mutual Information Estimators. URL: <https://CRAN.R-project.org/package=mpmi> (дата обращения: 20.04.2026).
21. Kraskov A., Stogbauer H., Grassberger P. Estimating mutual information // *Physical Review E – Statistical, Nonlinear, and Soft Matter Physics*. 2004. vol. 69. no. 6. 066138 p. DOI: 10.1103/PhysRevE.69.066138.
22. Ross B.C. Mutual Information between Discrete and Continuous Data Sets // *PloS ONE*. 2014. vol. 9. no. 2. e87357 p. DOI: 10.1371/journal.pone.0087357.
23. Knnmi: k-Nearest Neighbor Mutual Information Estimator. URL: <https://CRAN.R-project.org/package=knnmi> (дата обращения: 20.04.2026).
24. YEAB: Analyze Data from Analysis of Behavior Experiments. URL: <https://CRAN.R-project.org/package=YEAB> (дата обращения: 20.04.2026).
25. Fan J., Song R. Sure Independence screening in generalized linear models with NP-dimensionality // *Annals of Statistics*. 2010. vol. 38. no. 6. pp. 3567–3604. DOI: 10.1214/10-AOS798.
26. Zyma I., Tukaev S., Seleznev I., et al. Electroencephalograms during Mental Arithmetic Task Performance // *Data*. 2019. vol. 4. no. 1. pp. 14. DOI: 10.3390/data4010014.
27. Timofeeva A.Y., Avdeenko T.V., Alkov S.S. A Study of Robust Partial Correlation between Electroencephalogram Connectivity and Human Arithmetic Abilities // *Pattern Recognition and Image Analysis*. 2025. vol. 35. pp. 189–200. DOI: 10.1134/S1054661825700105.
28. Zakharov I., Tabueva A., Adamovich T., et al. Alpha Band Resting-State EEG Connectivity is Associated with Non-verbal Intelligence // *Frontiers in Human Neuroscience*. 2020. vol. 14. no. 10. DOI: 10.3389/fnhum.2020.00010.
29. Авдеевко Т.В., Тимофеева А.Ю., Муртазина М.Ш., Разумникова О.М. Выявление устойчивых связей между показателями коннективности ЭЭГ и компонентами интеллекта // *Вестник Южно-Уральского государственного университета*. 2022. Т. 11. №4. С. 19–36.

Тимофеева Анастасия Юрьевна — канд. экон. наук, доцент, кафедра теоретической и прикладной информатики, Новосибирский государственный технический университет (НГТУ). Область научных интересов: развитие методов проектирования признаков для построения интерпретируемых моделей машинного обучения. Число научных публикаций — 100. a.timofeeva@corp.nstu.ru; пр. Карла Маркса, 20, 630073, Новосибирск, Россия; р.т.: +7(383)346-0600.

Авдеевко Татьяна Владимировна — д-р техн. наук, профессор, кафедра теоретической и прикладной информатики, Новосибирский государственный технический университет (НГТУ). Область научных интересов: моделирование и идентификация, математические и информационные методы поддержки принятия решений, инженерия знаний. Число научных публикаций — 250. avdeenko2101@gmail.com; пр. Карла Маркса, 20, 630073, Новосибирск, Россия; р.т.: +7(383)346-0600.

A. TIMOFEEVA, T. AVDEENKO

FEATURE SELECTION UNDER CLASS IMBALANCE: AN EMPIRICAL COMPARISON OF CORRELATION AND MUTUAL INFORMATION

Timofeeva A., Avdeenko T. Feature Selection under Class Imbalance: An Empirical Comparison of Correlation and Mutual Information.

Abstract. The problem of data imbalance in feature prescreening tasks using sure independence screening is considered. Within this framework, the choice of the dependence measure used for feature ranking is analyzed. This study investigates two of the most common association measures – the correlation coefficient and mutual information. The correlation coefficient characterizes linear dependence between variables; its estimates are computationally simple and possess favorable statistical properties. Mutual information is a more universal measure capable of detecting nonlinear dependencies; however, its estimation requires the reconstruction of distributions from sample data, which leads to increased computational complexity and sensitivity to parameter selection. Based on a series of computational experiments, it is shown that mutual information estimation methods based on k-nearest neighbors are inferior to kernel-based methods in terms of stability and screening quality. It is established that across all considered scenarios, including various levels of data imbalance, correlation-based measures provide more stable estimates and more reliable feature ranking compared to mutual information. The practical applicability of the results is demonstrated on real electroencephalography (EEG) data related to the analysis of arithmetic abilities. It is shown that a two-step procedure comprising screening based on the correlation coefficient followed by modeling using LASSO regression provides a substantial reduction in cross-validation error compared to approaches based on mutual information and those without prescreening. The use of the principal component constructed from the selected features further improves prediction accuracy. The obtained results indicate that in the context of small sample sizes, multicollinearity, and data imbalance, correlation-based measures are preferable for feature prescreening tasks in terms of stability and reproducibility.

Keywords: imbalanced data, mutual information, correlation coefficient, sure independence screening, kernel estimates, k-nearest neighbors, LASSO regression.

References

1. Liu J.Y., Zhong W., Li R.Z. A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics*. 2015. vol. 58. pp. 1–22. DOI: 10.1007/s11425-015-5062-9.
2. Jiang J., Shang J. Feature Screening for High-Dimensional Variable Selection in Generalized Linear Models. *Entropy*. 2023. vol. 25. no. 6. 851 p. DOI: 10.3390/e25060851.
3. Wasikowski M., Chen X. Combating the Small Sample Class Imbalance Problem Using Feature Selection. *IEEE Transactions on Knowledge and Data Engineering*. 2010. vol. 22. no. 10. pp. 1388–1400. DOI: 10.1109/TKDE.2009.187.
4. Chan J.Y.-L., Leow S.M.H., Bea K.T., et al. Mitigating the Multicollinearity Problem and its Machine Learning Approach: A Review. *Mathematics*. 2022. vol. 10. no. 8. 1283 p. DOI: 10.3390/math10081283.
5. Fan J., Fan Y. High Dimensional Classification Using Features Annealed Independence Rules. *Annals of Statistics*. 2008. vol. 36. no. 6. pp. 2605–2637. DOI: 10.1214/07-AOS504.

6. Xu Y., Qian Q. i-SISSO: Mutual information-based improved sure independent screening and sparsifying operator algorithm. *Engineering Applications of Artificial Intelligence*. 2022. vol. 116. 105442 p. DOI: 10.1016/j.engappai.2022.105442.
7. Shukla G., Singh S., Dhule C., et al. Point biserial correlation symbiotic organism search nanoengineering based drug delivery for tumor diagnosis. *Scientific Reports*. 2024. vol. 14. 6530 p. DOI: 10.1038/s41598-024-55159-6.
8. Huyen C. *Designing Machine Learning Systems*. Sebastopol: O'ReillyMedia. 2022. 443 p.
9. Mqadi N., Naicker N., Adeliyi T. Solving Misclassification of the Credit Card Imbalance Problem Using Near Miss. *Mathematical Problems in Engineering*. 2021. no. 1. 1–16 pp. DOI: 10.1155/2021/7194728.
10. Tasci E., Zhuge Y., Camphausen K., Krauze, A.V. Bias and Class Imbalance in Oncologic Data – Towards Inclusive and Transferrable AI in Large Scale Oncology Data Sets. *Cancers*. 2022. vol. 14. no. 12. 2897 p. DOI: 10.3390/cancers14122897.
11. Wongvorachan T., He S., Bulut O. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information*. 2023. vol. 14. no. 1. 54 p. DOI: 10.3390/info14010054.
12. Kumar A., Goel S., Sinha N., Bhardwaj A. A Review on Unbalanced Data Classification. *Proceedings of International Joint Conference on Advances in Computational Intelligence (IJCACI 2021)*. 2022. pp. 197–208. DOI: 10.1007/978-981-19-0332-8_14.
13. Mladenic D., Grobelnik M. Feature Selection for Unbalanced Class Distribution and Naive Bayes. *Proceedings of the Sixteenth International Conference on Machine Learning (ICML '99)*. 1999. pp. 258–267.
14. Wang L., Han M., Li X., et al. Review of Classification Methods on Unbalanced Data Sets. *IEEE Access*. 2021. vol. 9. pp. 64606–64628. DOI: 10.1109/ACCESS.2021.3074243.
15. Belghazi M.I., Baratin A., Rajeshwar S., et al. Mutual Information Neural Estimation. *Proceedings of the 35th International Conference on Machine Learning (PMLR)*. 2018. vol. 80. pp. 531–540.
16. Johnson N.L., Kots S., Balakrishnan N. *Odnomernyy'e nepreryvny'e raspredeleniya: ch. 2 [One-dimensional continuous distributions: part 2] (Per. 2-go angl. izd. Volkovoj O.I., Strigunovoj M.S., Shixovoj N.A. / Russ. ed.: Chepurina E.V). M.: BINOM. Laboratoriya znaniy. 2012. 600 p. (In Russ.)*.
17. Cheng Y., Liu H. A short note on the maximal point-biserial correlation under non-normality. *British Journal of Mathematical and Statistical Psychology*. 2016. vol. 69. no. 3. pp. 344–351. DOI: 10.1111/bmsp.12075.
18. Timofeeva A.Y. Data Type Detection for Choosing an Appropriate Correlation Coefficient in the Bivariate Case. *Proceedings of the Second Young Scientist's International Workshop on Trends in Information Processing (YSIP2)*. 2017. vol. 1837. pp. 188–194.
19. Czyz P., Grabowski F., Vogt J., Beerenwinkel N., Marx A. Beyond Normal: On the Evaluation of Mutual Information Estimators. *Advances in Neural Information Processing Systems*. 2023. vol. 36. pp. 1–34. DOI: 10.52202/075280-0742.
20. Mpmi: Mixed-Pair Mutual Information Estimators. Available at: <https://CRAN.R-project.org/package=mpmi> (accessed 20.04.2026).
21. Kraskov A., Stogbauer H., Grassberger P. Estimating mutual information. *Physical Review E – Statistical, Nonlinear, and Soft Matter Physics*. 2004. vol. 69. no. 6. 066138 p. DOI: 10.1103/PhysRevE.69.066138.
22. Ross B.C. Mutual Information between Discrete and Continuous Data Sets. *PloS ONE*. 2014. vol. 9. no. 2. e87357 p. DOI: 10.1371/journal.pone.0087357.
23. Knnmi: k-Nearest Neighbor Mutual Information Estimator. Available at: <https://CRAN.R-project.org/package=knnmi> (accessed 20.04.2026).
24. YEAB: Analyze Data from Analysis of Behavior Experiments. Available at: <https://CRAN.R-project.org/package=YEAB> (accessed 20.04.2026).

25. Fan J., Song R. Sure Independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics*. 2010. vol. 38. no. 6. pp. 3567–3604. DOI: 10.1214/10-AOS798.
26. Zyma I., Tukaev S., Seleznev I., et al. Electroencephalograms during Mental Arithmetic Task Performance. *Data*. 2019. vol. 4. no. 1. pp. 14. DOI: 10.3390/data4010014.
27. Timofeeva A.Y., Avdeenko T.V., Alkov S.S. A Study of Robust Partial Correlation between Electroencephalogram Connectivity and Human Arithmetic Abilities. *Pattern Recognition and Image Analysis*. 2025. vol. 35. pp. 189–200. DOI: 10.1134/S1054661825700105.
28. Zakharov I., Tabueva A., Adamovich T., et al. Alpha Band Resting-State EEG Connectivity is Associated with Non-verbal Intelligence. *Frontiers in Human Neuroscience*. 2020. vol. 14. no. 10. DOI: 10.3389/fnhum.2020.00010.
29. Avdeenko T.V., Timofeeva A.Y., Murtazina M.S., Razumnikova O.M. [Identification of stable relationships between EEG connectivity indicators and components of intelligence]. *Vestnik Yuzhno-Ural'skogo gosudarstvennogo universiteta – Bulletin of the South Ural State University*. 2022. vol. 11. no. 4. pp. 19–36. (In Russ.).

Timofeeva Anastasia — Ph.D., Associate Professor, Department of Theoretical and Applied Informatics, Novosibirsk State Technical University (NSTU). Research interests: the development of feature engineering methods for building interpretable machine learning models. The number of publications — 100. a.timofeeva@corp.nstu.ru; 20, Karl Marx Ave., 630073, Novosibirsk, Russia; office phone: +7(383)346-0600.

Avdeenko Tatyana — Ph.D., Dr.Sci., Professor, Department of Theoretical and Applied Informatics, Novosibirsk State Technical University (NSTU). Research interests: modeling and identification, mathematical and information methods for decision-making support, knowledge engineering. The number of publications — 250. avdeenko2101@gmail.com; 20, Karl Marx Ave., 630073, Novosibirsk, Russia; office phone: +7(383)346-0600.