

МЕТОДИКА БЫСТРОГО АВТОМАТИЗИРОВАННОГО ДИСКРИМИНАНТНОГО АНАЛИЗА ЭЭГ ПРИ РАЗЛИЧИИ В ДВА КЛАССА

А.А. МИТРОФАНОВ^а, И.В. КИЧУК^б, М.Н. РУСАЛОВА^с,
С.В. ЧАУСОВА^б, Н.В. СОЛОВЬЕВА^д

^а Федеральное государственное бюджетное научное учреждение «Научный центр психического здоровья», 115522, Россия, Москва, Каширское шоссе, д. 34

^б ГБОУ ВПО РНИМУ им Н.И. Пирогова Минздрава России, 117997, Россия, Москва, ул. Островитянова, д. 1

^с Институт высшей нервной деятельности и нейрофизиологии Российской академии наук, 117485, Россия, Москва, ул. Бутлерова, д. 5а

^д Акционерное общество «Научный центр персонализированной психиатрии», 105082, Россия, Москва, Большая Почтовая ул., д. 20, стр. 3

Резюме

Авторы предлагают методику ускоренных процедур дискриминантного анализа с использованием Программного комплекса анализа и топографического картирования электрической активности мозга «Brainsys» для проведения дискриминантного анализа ЭЭГ в два класса. Программный комплекс позволяет рассматривать до 10 тысяч спектральных и прочих параметров ЭЭГ, которые могут потенциально выступить в качестве предикторов линейной дискриминантной функции. Предварительно параметры ЭЭГ могут быть приведены к приблизительно нормальному распределению. Предложены различные процедуры отбора параметров ЭЭГ, чтобы они удовлетворяли модели Фишера, а также процедуры поиска наиболее информативных параметров – кандидатов в предикторы. Были рассмотрены процедуры пошагового включения предикторов и процедуры перебора всех возможных комбинаций предикторов. Процедуры включают в себя поиск по нескольким критериям оценки наилучшей дискриминации на обучающих выборках. Особое внимание было уделено проблеме переобучения. Завышенные оценки чувствительности и специфичности, полученные на обучающих выборках, могут возникнуть как из-за отклонений от модели Фишера, так и множественного тестирования в процедуре перебора всех возможных комбинаций. Решение проблемы переобучения следует искать в подтверждении оценки качества различения на независимых выборках, как минимум одной. Также рассмотрены вопросы получения оценки обобщающей способности (способность к правильному предсказанию на независимых тестовых выборках) линейной дискриминантной функции, полученной в результате дискриминантного анализа. Методика позволяет без рутинных расчетов проверить множество вариантов дискриминантных функций или получить быстрые предварительные оценочные результаты, применив процедуру пошагового включения предикторов. Была получена в качестве примера формула линейной дискриминантной функции, включающая спектральные параметры ЭЭГ, для различения лиц импульсивного поведения и склонных к самоконтролю.

Ключевые слова: импульсивность, самоконтроль, электроэнцефалограмма, дискриминантный анализ, обобщающая способность.

Дискриминантный анализ используется как для получения правила классификации многомерных случайных величин на основании имеющихся данных (о принадлежности уже полученных значений к тому или иному классу), так и для определения переменных, которые вносят наибольший вклад в различение классов. Для дискриминации в два класса часто применяют линейную дискриминантную функцию (ЛДФ) Фишера, которая требует, чтобы многомерные случайные величины имели нормальное распределение и, соответственно, нормальности распределения каждой независимой случайной величины (предиктора), а также равенства ковариационных матриц. В этом случае линейная дискриминантная функция для дискриминации в два класса представлена следующим выражением в матричном виде (Айвазян и др., 1989; Андерсон, 1963; Ledermann, Lloyd, 1984):

$$y = x^T C^{-1} (\bar{x}_1 - \bar{x}_2) - \frac{1}{2} (\bar{x}_1 + \bar{x}_2)^T C^{-1} (\bar{x}_1 - \bar{x}_2), \quad (1)$$

где

y – скалярная величина результатов вычислений (дискриминантные баллы); если $y \geq 0$, то новое наблюдение следует отнести к первому классу, в противном случае – ко второму;

x^T – вектор-строка нового наблюдения, которое нужно отнести к одному из двух классов;

C^{-1} – обратная матрица к объединенной выборочной ковариационной матрице;

\bar{x}_1 и \bar{x}_2 – векторы-столбцы выборочных средних предикторов;

T – знак транспонирования матрицы (вектора).

ЛДФ с p предикторами также можно представить в более удобном виде уравнения регрессии:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p, \quad (2)$$

где

y – скалярная величина результатов вычислений (дискриминантные баллы), причем, если $y \geq 0$, то наблюдение следует отнести к первому классу, если $y < 0$, то ко второму;

$x_1 \dots x_p$ – p значений многомерного наблюдения;

$b_1 \dots b_p$ – нестандартизованные коэффициенты ЛДФ;

b_0 – константа (свободный член).

Дискриминантные коэффициенты могут быть представлены в нестандартизованной, как в уравнении (2), и стандартизованной форме. Если нас интересует относительный вклад переменной, то коэффициенты следует представлять в стандартизованной форме (Z -преобразование), которые получаются из соотношения (1) при нормировании исходных переменных таким образом, чтобы математическое ожидание равнялось нулю, а стандартное отклонение равнялось единице (Клеца, 1980).

Для простоты изложения в дальнейшем полагаем, что цена ошибочной классификации одинакова для обоих классов.

Центроид — это точка с координатами в виде средних каждой многомерной переменной. Модель Фишера позволяет использовать (в частности, для теоретической оценки качества дискриминации) квадрат расстояния Махаланобиса D^2 (Mahalanobis Squared Distance) между центроидами классов, который учитывает коррелированность переменных, имеющих разные единицы измерения и дисперсии (Klecka, 1980; Айвазян и др., 1989). Расстояние Махаланобиса D^2 между центроидами многомерных выборок равно:

$$D^2 = (\bar{x}_1 - \bar{x}_2)^T C^{-1} (\bar{x}_1 - \bar{x}_2), \tag{3}$$

где

C^{-1} — матрица, обратная объединенной выборочной ковариационной матрице;

\bar{x}_1 и \bar{x}_2 — векторы-столбцы выборочных средних предикторов;

T — знак транспонирования матрицы (вектора).

Чем больше расстояние Махаланобиса между центроидами классов, тем выше различение. Если известны параметры нормального распределения многомерных случайных величин, то случайная величина u для наблюдений, относящихся к первому классу, имеет нормальное распределение с математическим ожиданием $D^2/2$ и стандартным отклонением D (дисперсией D^2), а случайная величина y для наблюдений, относящихся ко второму классу, имеет нормальное распределение с математическим ожиданием $-D^2/2$ и также стандартным отклонением D . При этом ошибка классификации Err (если цены ошибочной классификации равны) для обоих классов равна $\Phi(D/2)$ (Anderson, 1963), где Φ — стандартная функция нормального распределения. Однако важно отметить, что нам известны только выборочные оценки параметров распределения, поэтому оценка качества дискриминации, основанная на выборочной оценке квадрата расстояния Махаланобиса, будет приблизительной.

В таблице 1 приведены ошибки различения Err в процентах в зависимости от квадрата расстояния Махаланобиса D^2 . Таким образом, расстояние Махаланобиса дает представление о возможной ошибке классификации в идеальном случае, т.е. о ее нижней границе, которая, однако, может быть приблизительной в случае, например, значительного отклонения распределения исходных данных от нормального. Из таблицы 1 видно, что если расстояние Махаланобиса меньше 2, то ошибка разделения в два класса будет недопустимо большой и правило дискриминации следует признать неудачным.

Вероятность ошибки классификации нового наблюдения E (т.е. вероятность принадлежности к другому классу) можно оценить при помощи формулы

Таблица 1

Ошибка различения Err в процентах в зависимости от расстояния Махаланобиса D^2 между центроидами выборок в модели Фишера

D^2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$Err, \%$	31	24	19	16	13	11	9	8	6.7	5.7	4.9	4.1	3.6	3.1	2.6	2.3

вероятности отклонения от центроида при нормальном распределении (при условии, что цены ошибочной классификации равны между собой):

$$E = \Phi((|y| + D^2/2)/D),$$

где

Φ — стандартная функция нормального распределения;

y — дискриминантные баллы;

D — расстояние Махаланобиса между центроидами.

С геометрической точки зрения линейная дискриминантная функция является гиперплоскостью в r -мерном пространстве, которая делит пополам отрезок, соединяющий два центроида, и перпендикулярна ему (Kendall, Stuart, 1968). Она проведена оптимальным образом, так, чтобы минимизировать случаи ошибочной классификации (т.е. сделать минимальным количество случаев ошибочного отнесения к другому классу).

Задача дискриминации нового наблюдения объекта X сводится к проверке простой гипотезы H_0 « X принадлежит классу № 1» против простой альтернативной гипотезы H_1 « X принадлежит классу № 2». По итогам проверки на экзаменующей выборке получаем таблицу 2 (Айвазян и др., 1989).

Отметим следующие характеристики полученного правила (критерия) дискриминации объекта в класс № 1 (Там же):

1) Чувствительность равна доле случаев правильной классификации объектов класса № 1 $a/(a + e)$. Ошибка первого рода α в проверке гипотезы (вероятность, что гипотеза H_0 неправильно отвергнута, т.е. вероятность ложного отнесения к классу № 2) равна

$$\alpha = 1 - a/(a + e). \quad (4)$$

2) Специфичность равна доле случаев правильной классификации в класс № 2 $f/(b+f)$ и представляет собой мощность критерия дискриминации. Ошибка второго рода β (вероятность того, что H_0 неправильно принята, т.е. вероятность ложного отнесения к классу № 1) в проверке гипотезы равна

$$\beta = 1 - f/(b + f). \quad (5)$$

Для правила (критерия) дискриминации объекта в класс № 2 характеристики аналогичны. Заметим, что ошибка первого рода в правиле дискриминации в класс № 1 становится ошибкой второго рода в правиле дискриминации в класс № 2, а ошибка второго рода, соответственно, ошибкой первого рода, поэтому доли случаев правильной классификации для обоих классов, т.е. чув-

Таблица 2

Итоги проверки на экзаменующей выборке

Результат применения ЛДФ	Объект принадлежит классу № 1	Объект принадлежит классу № 2	Всего
Предсказан класс № 1 (принимается гипотеза H_0)	a	b	a + b
Предсказан класс № 2 (принимается гипотеза H_1)	e	f	e + f
Всего	a + e	b + f	N

ствительность и специфичность, обозначим как «результативность» правила дискриминации для данного класса.

В процессе дискриминантного анализа необходимо, как правило, получить ЛДФ с минимальной ошибкой первого рода α при допустимой мощности, т.е. достаточно малой ошибкой второго рода β , что эквивалентно получению ЛДФ с максимальной чувствительностью при допустимой специфичности.

Применение дискриминантного анализа при исследовании электрической активности мозга (количественных параметров электроэнцефалограммы — ЭЭГ) сталкивается с рядом сложностей, связанных, в первую очередь, с очень большим числом (до 10 тысяч и даже более) спектральных и прочих параметров ЭЭГ, которые могут потенциально выступить в качестве предикторов в ЛДФ. Для отбора наиболее информативных переменных — кандидатов в предикторы — используется ряд методов, которые, однако, не дают полной гарантии нахождения оптимального набора предикторов.

Также необходимо отметить, что при дискриминантном анализе ЭЭГ параметры многомерного нормального распределения показателей ЭЭГ неизвестны и можно получить только их выборочные оценки. В этих условиях теоретически вычислить вероятность ошибочной дискриминации очень сложно (Anderson, 1963), а результаты носят случайный характер. Кроме того, распределение показателей ЭЭГ необходимо привести к нормальному, при этом само распределение может быть нормальным только приблизительно.

Самая сложная проблема — выбор предикторов из множества независимых переменных, которых при исследовании ЭЭГ может быть несколько тысяч, что приводит к весьма длительным вычислениям, так как вычислительные мощности исследователя ограничены.

Допустим, что мы ищем дискриминантное правило для натурального логарифма мощности отведений стандартных ритмов дельта (0.5–4 Гц), тета (4–8 Гц), альфа (8–13 Гц), бета1 (13–20 Гц) и бета2 (20–30 Гц). Естественно включить в набор переменных в первую очередь те, у которых обнаружались значимые различия (Клеца, 1980). Однако переменные с незначимыми различиями, несмотря на предполагаемую малоинформативность, могут существенно улучшить ЛДФ (Айвазян и др., 1989), поэтому рационально по возможности включать все переменные, таким образом, число независимых переменных при 16 отведениях регистрации ЭЭГ будет равно 80.

Можно предложить различные процедуры отбора параметров ЭЭГ, например, отбирать параметры ЭЭГ с наибольшим уровнем значимости различий при сравнении по критерию Стьюдента (Лапин, Митрофанов, 2017). Самым популярным способом решения этой проблемы является использование метода пошагового включения предикторов (Клеца, 1980; Gevins, Remond, 1987).

Процедуру пошагового включения начинают с выбора наилучшей ЛДФ с одним предиктором и затем пошагово добавляют предиктор из множества исходных параметров ЭЭГ, который также дает наилучшее качество различения. Поиск прекращается, если качество различения перестало улучшаться.

В качестве меры различения (оценки качества) предложены разные величины: лямбда Уилкса, частное F-отношение, V-статистика Рао, межгрупповая

F-статистка, минимизация остаточной дисперсии, расстояние Махаланобиса (Klecka, 1980). При дискриминации в два класса, как в нашем случае, обычно рассматривают расстояние Махаланобиса между центроидами выборок; если оно перестает значимо увеличиваться при добавлении очередного предиктора, то поиск прекращают (Афифи, Эйзен, 1982).

Следует заметить, что расстояние Махаланобиса также неизвестно, а известна только его выборочная оценка, которая может весьма значительно отличаться от истинной. Можно предложить другую меру различения — сумму оценок чувствительности и специфичности по обучающим выборкам, максимум которой считается наилучшей дискриминацией, что эквивалентно минимуму суммы ошибок $\alpha + \beta$. У этой оценки есть недостаток — возможно, что чувствительность или специфичность будут слишком малы. В этом случае может оказаться полезной оценка в виде наименьшего значения из чувствительности и специфичности для каждого наблюдения из обучающей выборки — максимум этой оценки и будет означать наилучшую дискриминацию на данном шаге.

Возможно, что с практической точки зрения важна только максимальная чувствительность (или максимальная специфичность), тогда в качестве оценки качества различения можно указать только чувствительность (или только специфичность).

Таким образом, можно осуществлять процедуру пошагового включения для дискриминантного анализа в два класса по нескольким критериям оценки наилучшей дискриминации. С геометрической точки зрения мы ищем несколько гиперплоскостей, которые по-разному в зависимости от способа оценки качества разделяют две выборки.

Также может быть реализована процедура пошагового исключения, когда на каждом шаге выбрасывается самая малоинформативная переменная (Klecka, 1980). Первоначальное множество переменных, однако, не должно быть слишком велико, не более 200–300, чтобы вычисления не затягивались, да и точность вычислений при работе с обратными матрицами может не позволить работать с большим числом переменных.

Для решения проблемы выбора способа формирования подмножества наиболее информативных переменных можно предложить алгоритм, когда во множество переменных включаются все переменные, отобранные в процессе процедуры пошагового включения по нескольким критериям оценки наилучшей дискриминации.

Процедура пошагового отбора может не дать наилучшей комбинации предикторов. Чтобы найти наилучшее решение, необходимо проверить все возможные сочетания, начиная с 2, 3, 4 и т.д. предикторов, а этот алгоритм требует больших вычислительных и временных ресурсов (Ibid.). Если обычный компьютер за несколько дней может для нескольких тысяч исходных параметров ЭЭГ проверить все пары, тройки и даже четверки, то следующее увеличение числа переменных требует уже суперкомпьютеров.

Разные способы поиска могут компенсировать ошибки — выбросы, ошибки и артефакты при регистрации ЭЭГ или в случае, когда данные не удовлетворяют математическим допущениям дискриминантного анализа.

Существенно, что нарушения допущений дискриминантного анализа могут не носить критического характера (Айвазян и др., 1989). У. Клекка отмечал: «Исследователю, главный интерес которого состоит в математической модели, способной хорошо предсказывать или служить разумным описанием реального мира, самым лучшим руководством является процент правильных классификаций. Если этот процент высок, то нарушения математических допущений не наносит большого ущерба» (Ibid.). Таким образом, при анализе ЭЭГ можно довольно сильно смягчать условия модели Фишера, т.е. нормальность распределения и равенства ковариационных матриц, поскольку важен конечный результат, который, однако, требует подтверждения на независимых наборах данных.

Однако оценка качества различения только по обучающим выборкам будет завышенной, потому что известны не истинные, а только выборочные параметры распределения и дискриминантная функция отражает свойства конкретной выборки более точно, чем свойства всей генеральной совокупности (Клекка, 1980). Если целью дискриминантного анализа является дискриминация новых наблюдений с достаточной точностью, то необходимо провести процедуру подтверждения (validation) прогностической способности (prediction ability) полученной линейной дискриминантной функции хотя бы на одной новой независимой выборке.

Способность к обобщению, или обобщающая способность (generalization performance), обучающего метода (в том числе дискриминантного анализа) определяется как способность к правильному предсказанию на независимых тестовых выборках (Hastie et al., 2009). «Переобучение» (overfitting), или слишком оптимистическая оценка по обучающим выборкам качества различения (предсказательной способности), возникает, в частности, из-за высокой сложности модели, когда слишком много предикторов, тогда на обучающих выборках можно получить чрезмерно хорошо подогнанную модель, которая учитывает ошибки измерения, погрешности модели и пр. в обучающей выборке, которые отсутствуют в другой новой независимой выборке. Для всякой задачи существует оптимальная сложность модели, при которой достигается наилучшее качество обобщения (Vapnik, 1998; Воронцов, 2004). При линейном дискриминантном анализе ЭЭГ оптимальную сложность модели определить затруднительно, но практика в использовании процедуры пошагового включения предикторов показывает, что число предикторов может колебаться, как правило, от 2 до 5–6, в редких случаях превышать 6, дальнейшее усложнение модели увеличивает ошибку классификации уже при тестировании на самих обучающих выборках.

Важно отметить, что на самом деле есть две отдельные цели, которые преследует исследователь при проведении дискриминантного анализа (Hastie et al., 2009):

- 1) выбор модели: оценка эффективности различных моделей для того, чтобы выбрать самую лучшую;

- 2) оценка модели: выбор окончательной модели, оценка ее ошибки прогнозирования (ошибки обобщения, generalization error) по новым данным.

Первая цель позволяет выявить те переменные, которые вносят наибольший вклад в различение двух выборок. Вторая цель дает правило дискриминации в два класса, пригодное, чтобы предсказать принадлежность нового наблюдения к одному из классов с достаточно большой вероятностью. При ограниченном объеме данных можно достичь только первой цели, что касается второй цели, то на подтверждение и тестирование обобщающей способности может потребоваться выделение значительной доли материала, по разным оценкам, от одной трети до половины.

Помимо метода разделения исходной выборки на обучающую и подтверждающую, предложены другие различные методы оценки обобщающей способности. Приведем наиболее известные и простые с точки зрения реализации. Если данных совсем немного, то самый простой и широко используемый метод оценки ошибки предсказания — перекрестная проверка (cross-validation) (Hastie et al., 2009), в частности, скользящий экзамен (leave-one-out cross-validation, LOO CV), который в качестве экзаменующего наблюдения использует наблюдение, которое удаляется из исходной выборки, при этом ЛДФ пересчитывается вновь, после чего наблюдение возвращается в набор данных. Таким образом перебираются все исходные наблюдения и вычисляются ошибка предсказания и, соответственно, чувствительность и специфичность правила дискриминации. Оценка, полученная в результате скользящего экзамена, является несмещенной, но может иметь большую дисперсию (Ibid.). Однако скользящий экзамен не предотвращает эффекта переобучения из-за отклонения распределения переменных в данной выборке от распределения в генеральной совокупности, что подтвердилось в ходе различных исследований с применением данного программного обеспечения.

Если же мы имеем достаточно много исходных данных, то наилучшим подходом (Ibid.) будет разбиение случайным образом всего набора данных на обучающую, подтверждающую и тестирующую выборки, например, в пропорции 50%, 25%, 25%, хотя общепринятых рекомендаций по пропорциям нет. Обучающая выборка (training set) используется для получения ЛДФ; подтверждающая (validation set) — для оценки ошибки предсказания для выбранной модели; тестирующая (test set) — для оценки ошибки обобщающей способности. Тестирующая выборка должна использоваться только в самом конце анализа, иначе она покажет заниженную оценку ошибки обобщающей способности (generalization error) (Ibid.). Разбиение исходных данных на три выборки (обучающую, подтверждающую и тестирующую) близко к разбиению исходного материала только на две выборки (обучающую и подтверждающую), изложенную Клеккой (Klecka, 1980), и при достаточно большом объеме материала выглядит более предпочтительным. Что касается величины обучающей выборки, то трудно найти общее правило, сколько данных достаточно для дискриминации, это зависит, например, от сложности применяемой модели (Hastie et al., 2009) и других факторов.

Дискриминантный анализ имеет смысл проводить, если имеются значимые различия между средними многомерных выборок. Для выявления значимых различий между многомерными выборками может проводиться многомерный

однофакторный дисперсионный анализ с использованием статистики лямбда Уилкса.

Обычно для дискриминантного анализа используют статистические пакеты SPSS, Statistica и пр. В них предусмотрены различные процедуры удаления малоинформативных или включение информативных переменных в пошаговых процедурах. Однако сам процесс загрузки исходных данных зачастую весьма трудоемок, так как исходные данные необходимо подготовить и загрузить в статистические пакеты, а также перебирать возможные варианты с различными параметрами по эвристическим алгоритмам.

В настоящей работе предложены другие подходы. В компьютерный Программный комплекс топографического картирования электрической активности мозга «Brainsys» включены программные средства, разработанные при финансировании ООО «Научно-медицинская фирма “Статокин”, в которых реализованы возможности пошагового включения/исключения предикторов, а также процедура перебора всех вариантов наборов предикторов (Клеска, 1980), как из всего множества независимых переменных, так и наиболее подходящих, например, тех, у которых уровень значимости различий достаточно мал и которые удовлетворяют условиям проведения дискриминантного анализа. При этом рутинная часть сильно сокращена и упрощена, исследователь не тратит время на подготовку данных в виде таблиц Excel, достаточно только указать папки, которые содержат обучающие, подтверждающие и тестовые выборки в виде ЭЭГ-файлов в формате EDF, метод и параметры, программа сама реализует процедуры дискриминантного анализа. Представленная компьютерная программа дискриминантного анализа может использоваться как самостоятельно, так и в составе Комплекса аппаратно-программного для топографического картирования электрической активности мозга «Нейро-КМ», который представляет собой компьютерный электроэнцефалограф с мощным пакетом статистической обработки ЭЭГ и других биосигналов для научных исследований.

Для дискриминантного анализа могут отбираться те параметры ЭЭГ, которые не имеют значимых различий дисперсии по критерию Фишера ($p < 0.05$ или $p < 0.01$ — на усмотрение исследователя), а также приближены к распределению Гаусса, хотя эти требования не являются критичными и допускаются некоторые отклонения (Ibid.). Для каждого из отведений стандартной системы отведений 10–20 для частотных полос с шагом 1, 1.5 и 2 Гц и для стандартных частотных диапазонов (дельта, тета, альфа, бета1 и бета2) можно рассчитать и включить в исходный набор параметров следующие параметры ЭЭГ, имеющие приблизительно нормальное распределение (Бендат, Пирсол, 1974; Gevins, Remond, 1987; Нюер, 1992):

1) натуральный логарифм мощности Power на заданной частотной полосе, сокращенно $\text{Ln}(P)$:

$$\text{Ln}(P) = \text{Ln}(\text{Power}),$$

где Power — спектральная мощность на данном отведении и частотной полосе (пример: $(\text{Ln}(P)(O2)8-13)$ — натуральный логарифм спектральной мощности альфа-ритма (8–13 Гц) в отведении O2);

2) нормализованная (приведенная к нормальному распределению) относительная мощность LRP, вычисляемая по формуле:

$$LRP = \text{Ln}(\text{Power}(\%)/(100 - \text{Power}(\%))),$$

где Power (%) – относительная мощность данного стандартного ритма в процентах от мощности всей частотной полосы 0.5–30 Гц (пример: (LRP(O2)8-13) – параметр, вычисляемый по формуле: $C = \text{Ln}(S/(100 - S))$, где S – отношение мощности альфа-ритма (8–13 Гц) к суммарной мощности биопотенциалов в диапазоне 0.5–30 Гц в процентах на отведении F8);

3) нормализованная когерентность LCoh:

$$LCoh = \text{Ln}(\text{Coh}^2/(1 - \text{Coh}^2)),$$

где Coh^2 – квадрат модуля когерентности (КМК) (пример: LCoh(F3-Fp1)0.5-4 – параметр, вычисляемый по формуле: $LCoh = \text{Ln}(\text{Coh}^2/(100 - \text{Coh}^2))$, где Coh^2 – квадрат модуля когерентности дельта-ритма (0.5–4 Гц) между отведениями F3-Fp1);

4) межполушарная асимметрия мощности Asym:

$$Asym = \text{Ln}(\text{Power}(\text{Right})/\text{Power}(\text{Left})),$$

где Power(Right) и Power(Left) – мощность справа и слева соответственно для заданного отведения и частотной полосы (пример: Asym(O2/O1)4-8 – логарифм отношения мощности тета-ритма на отведении O2 к мощности тета-ритма на симметричном отведении O1);

5) натуральный логарифм канонограмм стандартных ритмов $\text{Ln}(P(B1)/P(B2))$, где P(B1) и P(B2) – мощность на частотных полосах B1 и B2 соответственно (пример: $(\text{Ln}(P(8-13)/P(0.5-4)))T4$) – натуральный логарифм отношения спектральной мощности альфа-ритма к спектральной мощности дельта-ритма в отведении T4).

Также можно включать и другие характеристики ЭЭГ, приведенные к нормальному распределению, например, действительную и мнимую части когерентности, их асимметрию, отношение мощностей ритмов с шагом 1 Гц и т.д.

Нормальность распределения также при необходимости можно проверить по критерию Колмогорова–Смирнова и удалить переменные, которые ему не удовлетворяют.

Процедура перебора вариантов заключается в том, что после предварительного отбора по тем или иным правилам независимых переменных, среди которых будем искать предикторы, проверяем все возможные их сочетания, начиная с 2, 3, 4 и т.д. (Клеца, 1980), и таким образом находим наилучшие решения. Полученные линейные дискриминантные функции могут тестироваться на обучающей выборке (в необходимых случаях) методом скользящего экзамена, на подтверждающих, тестовых и других контрольных выборках (Митрофанов и др., 2019).

В процессе работы те результаты, ошибка которых не превышает заданный уровень значимости α и β для экзаменационной и обучающей выборок, выводятся в таблицу Excel. Если же удовлетворительных результатов не найдено, то выводится максимально правильный результат с точки зрения различных критериев поиска (квадрат расстояния Махаланобиса, минимальная сумма ошибок α и β и пр.).

Если ставится цель оценки обобщающей способности и если позволяет количество исходных данных, то возможно по итогам анализа протестировать полученные ЛДФ на тестовых выборках.

ЛДФ также можно включить в виде экспертной системы в состав программного обеспечения и проводить дискриминацию как отдельных файлов ЭЭГ-записей, так и всех ЭЭГ-файлов в отдельных папках с выводом результатов классификации и описательной статистики в текстовой файл.

Время вычислений в процедуре перебора всех возможных комбинаций предикторов линейно зависит от объемов экзаменуемых выборок, в большей степени – от числа предикторов n и количества исходных независимых переменных m , практически оно пропорционально числу сочетаний из n предикторов по m независимым переменным C_m^n . Допустим, для поиска ЛДФ методом перебора всех вариантов для четырех предикторов на исходном множестве в 80 независимых переменных (если рассматривать 16 отведений для пяти стандартных ритмов) при объеме экзаменуемой выборки, равном 10, необходимо перебрать 1 581 580 вариантов, на что уходит примерно до минуты на довольно среднем по производительности компьютере, для пяти предикторов из 80, соответственно, около 10 минут (24 млн вариантов), для 6 – примерно 2 часа (300 млн) и т.д.

В рассматриваемом программном комплексе число исходных независимых переменных может быть до 10 000. Если число независимых переменных близко к максимуму, т.е. около 10 000, то для перебора всех вариантов ЛДФ с двумя предикторами может потребоваться несколько часов, для ЛДФ с тремя предикторами – несколько суток и т.д. В некоторых случаях, когда для хорошего различения достаточно двух-трех предикторов, это вполне допустимо. Однако в большинстве случаев следует отобрать несколько сот наиболее подходящих кандидатов в предикторы дискриминантных функций, среди которых уже можно осуществлять поиск методом перебора. В процессе вычислений для отбрасывания заведомо неудачных и повторяющихся вариантов, а также малоинформативных переменных в программном комплексе «Brainsys» реализованы следующие процедуры:

1) исключение независимых переменных, у которых незначимо различаются средние, так как вероятность того, что они малоинформативны, выше, чем у переменных со значимыми различиями (Айвазян и др., 1989);

2) исключение независимых переменных, у которых в одномерном случае, т.е. при ЛДФ с одним предиктором, процент правильно угаданных наблюдений (чувствительность) довольно мал, например, ниже 50%;

3) анализ на выбросы среди наблюдений по F-критерию (Афифи, Эйзен, 1982);

4) анализ равенства средних многомерных нормальных выборок по уровню значимости лямбда Уилкса (Рао, 1973) или по критерию Хотеллинга (Афифи, Эйзен, 1982) с отбрасыванием вариантов комбинаций переменных с незначимыми различиями;

5) анализ множественного коэффициента корреляции для выявления переменных, которые являются линейной комбинацией других переменных;

6) F-удаление или F-включение малоинформативных переменных (Рао, 1973; Афифи, Эйзен, 1982);

7) возможность исключения из рассмотрения набора предикторов, если он уже включает набор предикторов успешных ЛДФ на предыдущих шагах.

Помимо метода перебора всех вариантов в программном комплексе «Brainsys» реализована упомянутая процедура пошагового включения предикторов, т.е. последовательного отбора наиболее полезных переменных (Klecka, 1980). Процедура пошагового включения предикторов позволяет очень быстро (за минуты) по сравнению с процедурой перебора всех вариантов получить результат, при этом исходный набор независимых переменных может быть размером до 10 000, но этот результат может оказаться не самым лучшим. Тем не менее этот метод весьма эффективен в ряде случаев, позволяет быстро получить удовлетворительный результат и оценить перспективность применения дискриминантного анализа для различения двух данных классов. В исходный набор можно включать практически все возможные параметры ЭЭГ, имеющие приблизительно нормальное распределение и не имеющие значимых различий в дисперсии по критерию Фишера.

В программном комплексе также реализована процедура дискриминации в три класса, которая может быть проведена, если известны ЛДФ для всех возможных пар классов (Ledermann, Lloyd, 1984). Пусть для трех классов А, В, С известны ЛДФ(А,В), ЛДФ(В,С) и ЛДФ(А,С). Новое наблюдение относим к классу А, если ЛДФ(А,В) и ЛДФ(А,С) одновременно отнесли это наблюдение к классу А.

Однако закономерно возникает вопрос, насколько используемые методики и ПО адекватны, не могут ли они дать ложноположительные или ложноотрицательные результаты. Для ответа на этот вопрос были проведены тестовые испытания на однородных выборках из реальных ЭЭГ лиц без зафиксированных нервно-психических расстройств.

Для тестирования были отобраны случайным образом две выборки ЭЭГ, каждая по 100 человек без нервно-психических отклонений, имитирующие два разных класса. Каждая выборка была, в свою очередь, согласно рекомендациям Клекки (Klecka, 1980), разбита на две выборки по 50 человек — обучающую и подтверждающую. Цель эксперимента на реальных данных состояла в проверке, не даст ли эта методика ложноположительного результата в виде ЛДФ с высокой чувствительностью и специфичностью. Для пошаговой процедуры по вышеперечисленным критериям были отобраны 10 000 параметров ЭЭГ с самыми большими различиями в убывающем порядке. Ни один из критериев поиска не дал ложноположительного результата с высокой чувствительностью и специфичностью, значения которых не превышали 66%, но при этом следует отметить, что процент правильно угаданных случаев на обучающих выборках рос по мере усложнения модели до 90%, но подтверждающие выборки давали реальный, т.е. очень низкий, результат, что являлось эффективным фильтром, отсекающим ЛДФ с переобучением. Далее был использован метод перебора всех возможных комбинаций, который показал аналогичные результаты, т.е. не было найдено правило различения двух этих

выборки с высокими значениями чувствительности и специфичности. Таким образом, подтверждено, что даже методика разбиения исходных данных на два набора не приведет к ложноположительным результатам в случае сравнения двух классов с одинаковым распределением. Следует ожидать, что при использовании различных усложненных алгоритмов дополнительную гарантию отсутствия переобучения может дать разбиение данных на три набора – обучающий, подтверждающий и тестирующий.

Также предусмотрена возможность имитационного моделирования данных. Имитационное моделирование (симулирование) исходных данных позволяет дополнительно верифицировать работоспособность алгоритмов и ПО, предоставляет пользователю возможность оценить качество алгоритмов и оптимальность сложности модели, а также в исследовательских и обучающих целях. Данные симулируются в виде многомерных нормально распределенных случайных величин, принадлежащим двум различным классам, причем в двух режимах. Первый режим симулирует произвольные нормально распределенные многомерные случайные величины с заданным расстоянием Махаланобиса. Второй режим позволяет симулировать нормально распределенные переменные с реальными параметрами распределения (средним и стандартным отклонением) показателей ЭЭГ. При этом учитываются корреляционные связи: в первом режиме линейные корреляции между переменными задаются случайным образом, во втором режиме берутся выборочные оценки коэффициента корреляции между реальными данными.

Рассмотрим пример дискриминации лиц с самоконтролем и импульсивных. Линейная дискриминантная формула (ЛДФ) была получена в результате дискриминантного анализа двух выборок ЭЭГ здоровых испытуемых от 17 до 45 лет. Для получения формулы были использованы записи ЭЭГ 104 испытуемых (из них 61 импульсивных и 43 лиц с самоконтролем, верифицированных на свойство импульсивность/самоконтроль при помощи психологических тестов). У всех испытуемых регистрировали ЭЭГ на установке, состоящей из 21-канального усилителя аппаратно-программного комплекса для топографического картирования электрической активности мозга «Нейро-КМ» (Научно-медицинской фирмы «Статокин») и персонального компьютера. Регистрацию биопотенциалов осуществляли по международной схеме 10–20% от 16 отведений: Fp1, Fp2, F3, F4, F7, F8, C3, C4, P3, P4, T3, T4, T5, T6, O1, O2. В качестве референтного использовали объединенный ушной электрод. Спектральный анализ ЭЭГ проводили с помощью программных средств АПК «Нейро-КМ». Эпоха анализа ЭЭГ составляла 1 с, частота опроса – 200 Гц. Полоса пропускания – 0.5–30 Гц.

Обработка данных ЭЭГ состояла в расчете спектров мощности и когерентности ЭЭГ в стандартной полосе частот (дельта, тета, альфа, бета1, бета2), а также сопоставлении их по процентному содержанию от всего диапазона частот.

Перед дискриминантным анализом полезно оценить, есть ли значимые различия между выборками. Картирование уровня значимости различий мощности стандартных ритмов обучающих выборок по критерию Стьюдента

(рисунок 1) показало, что выборочные средние логарифма мощности группы «Самоконтроль» превышают выборочные средние логарифма мощность группы «Импульсивность» по всем отведениям альфа и большинству отведений бета1 даже после применения весьма жесткой поправки Бонферрони на множественность сравнений. Столь большие различия делают весьма вероятной успешность решения задачи дискриминации, т.е. классификации многомерных случайных величин на основании уже имеющихся данных.

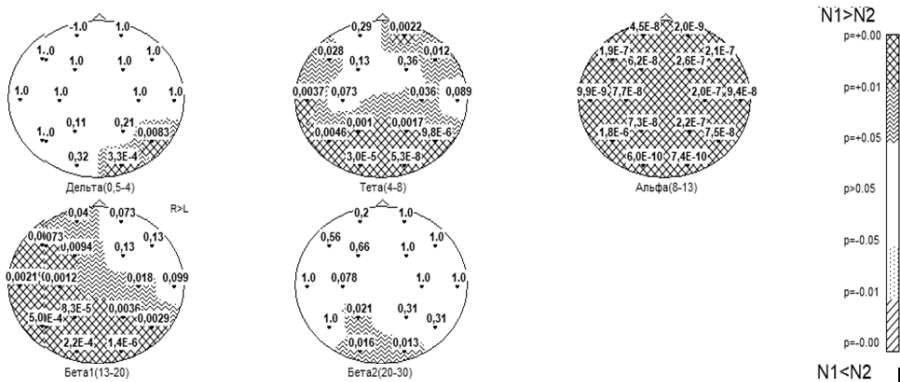
Частотные полосы 1 Гц, 1.5 Гц и 2 Гц не рассматривались, так как картирование показало большие значимые различия уже для стандартных ритмов.

Алгоритм действий при дискриминантном анализе следующий. Так как исходных данных немного, то воспользуемся рекомендацией Клекка (Клеска 1980) и разобьем случайным образом выборки на обучающую и экзаменуемую (подтверждающую). Из исходных ЭЭГ-файлов, очищенных от артефактов, случайным образом формируются обучающие выборки размером 22 человека для первого класса «Самоконтроль» и 31 человек для второго класса «Импульсивность», а также две экзаменуемые выборки для этих же классов в количестве 21 и 30 человек соответственно. Таким образом, на экзаменуемые выборки выделено почти половина материала, так как именно тестирование на экзаменуемых выборках критически важно для оценки качества найденного правила дискриминации.

Сначала попробуем использовать процедуру пошагового включения предикторов, а затем сравним ее с процедурой перебора всех возможных комбинаций.

Рисунок 1

Сравнение средних значений логарифма мощности (Ln(Power)) двух независимых выборок ЭЭГ-файлов N1 и N2 по двухстороннему критерию Стьюдента с поправкой Бонферрони



Примечание. N1 — выборка из класса «Самоконтроль», N2 — выборка из класса «Импульсивность». Картируется уровень значимости p. Положительный знак уровня значимости указывает, что среднее выборки N1 больше среднего выборки N2. Отрицательный знак уровня значимости указывает, что среднее выборки N1 меньше среднего выборки N2.

Процедура пошагового включения предикторов ЛДФ включала в себя поиск по нескольким критериям оценки наилучшей дискриминации на обучающих выборках. Предварительно были отобраны все показатели ЭЭГ для стандартных ритмов, у которых дисперсии не имели значимых различий по критерию Фишера ($p < 0.01$) и которые имеют близкое к нормальному распределение. Было найдено 990 таких показателей. По завершении каждого шага результат проверялся скользящим экзаменом и на подтверждающей выборке. Была получена ЛДФ для трех предикторов с чувствительностью 95% и специфичностью 93% и расстоянием Махаланобиса между центроидами 10.8:

$$y = 4.1 \times (\text{Ln}(P)(O1)8-13) + 2.0 \times (\text{LCoh}(F3-Fp1)4-8) + 0.67 \times (\text{Ln}(P)(O2)8-13) - 25.1$$

При значении показателя $y \geq 0$ поведенческую реакцию испытуемого диагностируют как склонность к самоконтролю, а при значении показателя $y < 0$ — к импульсивному поведению.

$(\text{Ln}(P)(O1)8-13)$ — натуральный логарифм спектральной мощности альфа-ритма (8–13 Гц) в отведении O1;

$(\text{LCoh}(F3-Fp1)4-8)$ — нормализованная когерентность между отведениями F3-Fp1 тета-ритма (4–8 Гц);

$(\text{Ln}(P)(O2)8-13)$ — натуральный логарифм спектральной мощности альфа-ритма (8–13 Гц) в отведении O1.

Уровень значимости для лямбда Уилкса $p < 5.4E-14$. Стандартизованные коэффициенты ЛДФ для различения в классы «Самоконтроль» и «Импульсивность», полученной процедурой пошагового включения (таблица 3), показывают вклад каждого предиктора. Теоретическая оценка ошибки классификации по выборочному значению расстояния Махаланобиса (таблица 1) не противоречит полученным результатам.

Рассмотрим вариант поиска ЛДФ при помощи процедуры перебора всех возможных комбинаций переменных.

Предварительная подготовка данных включала в себя поиск наиболее подходящих независимых переменных со значимостью различий между средними $p < 0.01$ по критерию Стьюдента (чтобы сократить число возможных комбинаций), а также значимостью различий между дисперсиями $p < 0.01$ по критерию Фишера.

В программе есть возможность поиска параметров, имеющих наибольшие различия, из которых и составляется упорядоченный по возрастанию уровня

Таблица 3

Стандартизованные коэффициенты ЛДФ для различения в классы «Самоконтроль» и «Импульсивность», полученной процедурой пошагового включения

Параметр	Стандартизованные коэффициенты
$(\text{Ln}(P)(O1)8-13)$	0.51
$(\text{LCoh}(F3-Fp1)4-8)$	0.39
$(\text{Ln}(P)(O2)8-13)$	0.10

значимости список независимых переменных (кандидатов в предикторы) для дискриминантного анализа. Всего было найдено 268 независимых переменных с уровнем значимости различий $p < 0.01$ (таблица 4).

При поиске ЛДФ были заданы пороги 95%-й чувствительности и специфичности для обучающих и экзаменуемых выборок, удовлетворяющие условиям поиска, выводились в таблицу Excel. Также проводилась процедура скользящего экзамена.

Были найдены несколько ЛДФ, среди которых наилучший результат с чувствительностью и специфичностью 95% для скользящего экзамена и подтверждающих выборок обоих классов дала следующая ЛДФ:

$$y = (\text{Ln}(P)(O1)8-13) + 0.15 \times (\text{Ln}(P(8-13)/P(0.5-4))T4) - 0.61 \times (\text{LRP}(F8)20-30) - 6.1,$$

где

$(\text{Ln}(P)(O1)8-13)$ – натуральный логарифм спектральной мощности альфа-ритма в отведении O1;

$(\text{Ln}(P(8-13)/P(0.5-4))T4)$ – натуральный логарифм отношения спектральной мощности альфа-ритма к спектральной мощности дельта-ритма в отведении T4;

$(\text{LRP}(F8)20-30)$ – параметр, вычисляемый по формуле: $C = \text{Ln}(S/(100 - S))$ где S – отношение мощности бета2-ритма к суммарной мощности биопотенциалов в диапазоне 0.5–30 Гц в процентах на отведении F8.

При значении показателя $y \geq 0$ поведенческую реакцию испытуемого диагностируют как склонность к самоконтролю, а при значении показателя $y < 0$ – к импульсивному поведению.

Уровень значимости для лямбда Уилкса $p < 1.1E-11$. Расстояние Махаланобиса равно 7.85, при этом значении теоретическая ошибка классификации равна примерно 8% (таблица 1), т.е. теоретически чувствительность и специфичность не могут превышать 92%, если не учитывать факт, что это выборочная оценка. Стандартизованные коэффициенты ЛДФ (таблица 5) свидетельствуют

Таблица 4

Уровень значимости наиболее отличающихся параметров по критерию Стьюдента при сравнении группы

№	Ритм	Параметр	Отведение	p
1	8.0–13.0 Гц	Ln(P)	(O1-AA)	7.5E-12
2	8.0–13.0 Гц	Ln(P)	(O2-AA)	9.2E-12
3	8.0–13.0 Гц	Ln(P)	(Fp2-AA)	2.5E-11
4	–	Ln(P(8-13)/P(0.5-4))	(Fp2-AA)	5.6E-11
5	–	Ln(P(8-13)/P(0.5-4))	(F3-AA)	1.1E-10
6	8.0–13.0 Гц	Ln(P)	(T3-AA)	1.2E-10
7	–	Ln(P(8-13)/P(0.5-4))	(Fp1-AA)	1.2E-10
8	8.0–13.0 Гц	Ln(P(8-13)/P(0.5-4))	(F4-AA)	3.9E-10
9	20.0–30.0 Гц	Ln(P)	(Fp1-AA)	5,6E-10
10	–	Ln(P(8-13)/P(0.5-4))	(C3-AA)	6.4E-10
...
267	4.0–8.0 Гц	LCoh	(C3-F7)	0.0095
268	20.0–30.0 Гц	LCoh	(F3-T5)	0.0098

о достаточно хорошей информативности предикторов. Наибольший вклад вносит натуральный логарифм спектральной мощности альфа-ритма в отведении О1.

Дальнейшее увеличение числа предикторов до 4 и более результаты не улучшило.

Таким образом, объем материала позволил путем разбиения исходных данных на обучающие и подтверждающие выборки и использования метода пошагового включения предикторов достичь первой цели – выбора наилучшей модели для данной выборки, выявить параметры ЭЭГ, которые в наибольшей степени вносят вклад в различение импульсивных и склонных к самоконтролю лиц, получить приблизительную оценку качества различения, но не позволил достичь второй цели – получить оценку обобщающей способности ЛДФ на тестовой выборке, которую, однако, нетрудно провести на новых данных.

Однако интересующий исследователя список параметров ЭЭГ, вносящих вклад в различение, не исчерпывается приведенными в найденной ЛДФ. При использовании процедуры перебора всех возможных комбинаций предикторов при заданных порогах 95%-й чувствительности и специфичности для обучающих и экзаменуемых выборок было найдено 229 ЛДФ, в которых фигурирует 141 предиктор как со знаком «плюс», так и со знаком «минус». В таблице 6 представлены наиболее часто встречающиеся предикторы (более чем 1%) в обнаруженных ЛДФ, что свидетельствует об их информативности при отнесении к тому или иному классу. Знак «плюс» показывает, что увеличение данного параметра увеличивает вероятность отнесения к группе «Самоконтроль», а знак «минус» – что увеличение данного параметра увеличивает вероятность отнесения к группе «Импульсивность». Следует ожидать, что чем чаще встречается предиктор в формулах ЛДФ, тем более значительную роль он играет в принадлежности к одному из двух классов.

Как видно из таблицы 6, для самоконтрольных лиц наиболее часто встречаемыми предикторами являются:

- 1) $LCoh(T3-C3)_{0.5-4}$ (68 случаев) – нормализованная когерентность между отведениями Т3-С3 в дельта-диапазоне;
- 2) $Ln(P)(O1)_{8-13}$ (52 случая) – логарифм мощности альфа на отведении О1;
- 3) $Ln(P)(O2)_{8-13}$ (49 случаев) – логарифм мощности альфа на отведении О2;

Таблица 5

**Стандартизованные коэффициенты ЛДФ для различения в классы
«Самоконтроль» и «Импульсивность»**

Параметр	Стандартизованные коэффициенты
$(Ln(P)(O1)_{8-13})$	0.61
$(Ln(P(8-13)/P(0.5-4))_{T4})$	0.09
$(LRP(F8)_{20-30})$	-0.31

Таблица 6

Частота встречаемости предикторов в ЛДФ, найденных методом перебора всех возможных комбинаций, для дискриминации в классы «Самоконтроль» и «Импульсивность» с чувствительностью и специфичностью не менее 95%

№	Предиктор	Найдено	%	Самоконтроль (со знаком «+»)	Импульсивность (со знаком «-»)
1	LCoh(T3-C3)0.5-4	68	29.7%	68	0
2	Ln(P)(O1)8-13	52	22.7%	52	0
3	Ln(P)(O2)8-13	49	21.4%	49	0
4	LCoh(F3-Fp1)0.5-4	44	19.2%	44	0
5	LRP(F8)20-30	37	16.2%	0	37
6	Ln(P)(T6)4-8	36	15.7%	36	0
7	LCoh(F3-Fp1)4-8	20	8.7%	20	0
8	Ln(P)(O2)13-20	19	8.3%	19	0
9	Ln(P)(T3)8-13	17	7.4%	17	0
10	Ln(P)(C3)8-13	16	7.0%	16	0
11	Ln(P)(O2)4-8	16	7.0%	16	0
12	LRP(O1)13-20	16	7.0%	0	16
13	Ln(P)(C4)8-13	13	5.7%	13	0
14	Ln(P)(Fp2)8-13	13	5.7%	13	0
15	Ln(P)(O1)4-8	11	4.8%	11	0
16	Ln(P(8-13)/P(13-20))_O1	9	3.9%	9	0
17	Ln(P)(T6)8-13	9	3.9%	9	0
18	Ln(P(8-13)/P(0.5-4))_O2	7	3.1%	7	0
19	Ln(P)(T3)13-20	7	3.1%	7	0
20	Ln(P(8-13)/P(0.5-4))_F3	7	3.1%	7	0
21	LRP(O1)0.5-4	6	2.6%	0	6
22	LCoh(T6-F4)0.5-4	5	2.2%	0	5
23	Ln(P(Альфа)/ P(Бета1+Бета2))_O1	5	2.2%	5	0
24	Ln(P(Альфа)/ P(Бета1+Бета2))_T3	5	2.2%	5	0
25	Ln(P(8-13)/P(0.5-4))_O1	5	2.2%	5	0
26	Ln(P)(F4)8-13	5	2.2%	5	0
27	LRP(O1)8-13	4	1.7%	4	0

4) LCoh(F3-Fp1)0.5-4 (44 случая) — нормализованная когерентность между отведениями F3-Fp1 в дельта-диапазоне.

Для импульсивных лиц характерны следующие предикторы:

LRP(F8)20-30 (37 случаев) — нормализованная относительная мощность бета2 в отведении F8.

Таким образом, когерентность в дельта-диапазоне в передних и передне-височных отделах левого полушария, а также мощность в альфа-полосе в затылочных отделах обоих полушарий в большей мере характерны для лиц, склонных к самоконтролю, тогда как для импульсивных лиц — мощность высоких частот в передних отделах правого полушария.

Интересные результаты показало имитационное моделирование. При имитационном моделировании данных во втором режиме (т.е. симуляции многомерных нормально распределенных величин и реальных выборочных средних, дисперсиях и ковариационной матрице для обоих классов) дискриминантного анализа для различения импульсивных лиц и лиц с самоконтролем были получены результаты, аналогичные результатам, полученным на реальных данных, но с более высокими чувствительностью и специфичностью, что можно объяснить возможными отклонениями реальных данных от нормального распределения. При этом полученная формула ЛДФ немного хуже, чем формула, полученная на реальных данных, но очень хорошо (с чувствительностью 93%, специфичностью 87%) дискриминирует исходные данные, что наталкивает на мысль, что при помощи имитационного моделирования, зная только выборочные средние и дисперсии переменных, а также корреляционную матрицу, вполне можно получать работающие формулы ЛДФ.

Подводя итоги, можно предложить следующий порядок проведения дискриминантного анализа для двух классов с использованием программного комплекса «Brainsys».

1. В зависимости от цели исследования и объема исходных данных решается вопрос, на сколько выборок делятся исходные данные.

В худшем случае (при очень малом объеме данных) можно использовать только обучающие выборки и скользящий экзамен, но тогда мы получим результаты (переменные, которые больше всего влияют на различение и их число) только для данной выборки с оценкой качества при помощи только скользящего экзамена, а также оценить уровень значимости различий двух многомерных выборок по лямбда-критерию Уилкса (который эквивалентен для двух классов критерию Хотеллинга) и перспективность более углубленного дискриминантного анализа по расстоянию Махаланобиса, но никак не сможем оценить риск переобучения.

При достаточном объеме данных и простых алгоритмах можно разбить исходные выборки на два набора — обучающий и подтверждающий, т.е. выделить из исходных выборок подтверждающие выборки, которые дадут приблизительную оценку качества различения для генеральных совокупностей обоих классов, лучшую, чем только по скользящему экзамену.

Для корректной оценки обобщающей способности, т.е. оценки правильной классификации нового наблюдения, исходный массив данных делится на три

группы: обучающая выборка (примерно 50%), подтверждающая (25%) и тестовая (25%), причем тестовые группы используются для получения оценки обобщающей способности только в самом конце анализа. Если же объем данных очень большой, то излишек можно отправить в тестовую группу, так как слишком большие обучающие и подтверждающие выборки могут существенно замедлить вычисления без существенного увеличения точности.

2. Вначале проводится дискриминантный анализ с пошаговым включением предикторов. Вполне возможно, что будут получены вполне удовлетворительные результаты и анализ на этом шаге можно завершить. В противном случае, если чувствительность и специфичность достаточно велики (скажем, больше 75%) при величине расстояния Махаланобиса более 3–4, имеет смысл попробовать более длительную процедуру перебора всех возможных комбинаций. Также исследователя может интересовать, какие переменные еще могут участвовать в разделении классов. Пошаговый анализ также дает представление о сложности модели, т.е. оптимальном числе предикторов для различения двух классов.

3. Процедура перебора всех возможных комбинаций требует больших затрат времени, которые растут по экспоненте с увеличением числа независимых переменных — кандидатов в предикторы. Если сложность модели невелика и уравнение ЛДФ включает 2–3 предиктора, то возможен перебор до 10 тыс. переменных — кандидатов в предикторы. В противном случае необходимо отбирать предположительно наиболее информативные переменные, чтобы их число не превышало нескольких десятков. Например, можно отобрать для анализа первые 50–100 самых различающихся по критерию Стьюдента переменных. Можно искать только среди параметров мощности, или относительной мощности, или когерентности, исключив отдельные стандартные диапазоны или полосы 1, 1.5 и 2 Гц. Можно исключать и отдельные отведения ЭЭГ, а также использовать различные параметры поиска. Также в качестве исходного набора переменных можно использовать те переменные, которые были получены в результате пошагового поиска. При использовании сложных алгоритмов с получением, например, множества различных вариантов может понадобиться дополнительный независимый контрольный набор данных, если возникают опасения возникновения эффекта переобучения.

4. Процедуры пошагового исключения малоинформативных предикторов полезны, например, при оценке вклада переменных ЭЭГ в различение двух классов, в частности, какие отведения реально вносят вклад в различение при сравнении двух выборок по альфа-ритму, например, или вообще по мощности.

Обсуждение

Предложенные методики дискриминантного анализа могут включать перебор множества вариантов, что может привести к трудностям, аналогичным проблеме множественного тестирования при проверке статистических гипотез. С большой степенью вероятности это может вызвать эффект переобучения и, соответственно, ложные результаты. Вообще «вопрос о качестве

алгоритмов, синтезированных по конечным выборкам прецедентов, является фундаментальной проблемой теории обучаемых систем» (Воронцов, 2004, с. 5). В нашем случае решение следует искать в подтверждении оценки качества различения на независимых выборках, как минимум одной, назовем ее «подтверждающей». При достаточном объеме материала его следует, согласно рекомендациям (Hastie et al., 2009), разделить на обучающий, подтверждающий и тестирующий наборы, причем тестирующий набор следует использовать только в самом конце анализа. Не исключено, что в сложных случаях может понадобиться дополнительный независимый контрольный набор данных (или выделение из исходного материала еще одного, скажем, «экзаменующего»), если возникли опасения возникновения этой проблемы в процессе перебора всех возможных комбинаций предикторов или применения сложных методик.

Опыт работы с предложенным ПО показывает, что оно весьма удобно для оценки перспектив различения двух классов по всем или части показателей ЭЭГ уже при использовании процедуры пошагового включения предикторов. Если чувствительность и специфичность достаточно велики, скажем, более 70–75% при довольно больших значениях квадрата расстояния Махаланобиса (более 2–3), то имеет смысл продолжать более тщательный поиск методом перебора всех возможных комбинаций, но весьма вероятно, что процедура пошагового поиска уже может дать удовлетворительный результат.

Тестирование методик и ПО на однородных выборках и имитационное моделирование показало, что при использовании сравнительно простого пошагового алгоритма включения предикторов методика оценки качества правила дискриминации двух классов путем разбиения исходных данных на две выборки (обучающую и подтверждающую) достаточно эффективна и не дает ложноположительных или ложноотрицательных результатов, однако при использовании более сложных алгоритмов, в том числе и алгоритма перебора всех возможных комбинаций предикторов, необходимы дополнительные независимые контрольные наборы данных, чтобы избежать получения излишне оптимистичной оценки качества полученного правила различения.

В нашем случае результат дискриминантного анализа представлен в виде множества предикторов, вносящих наибольший вклад в различение двух обучающих выборок, и подтвержден на независимых выборках. Были также получены оценки качества дискриминации. Правило дискриминации не тестировалось на обобщающую способность из-за небольшого объема исходного материала.

Результаты исследования показали, что основными предикторами различия ЭЭГ импульсивных и самоконтрольных (рефлексивных) лиц являются параметры альфа-, дельта- и бета2-ритмов.

По мнению большинства исследователей, альфа-ритм является универсальным коммуникативным кодом, поэтому его роли в интегративной деятельности мозга посвящено значительное число работ (Базанова, Афтанас, 2007; Кирой и др., 1990; Костандов и др., 2012; Owen et al., 2006). Предполагается, что альфа-система обеспечивает избирательную модуляцию корковой активности путем

изменения пространственно-временной организации ЭЭГ, благодаря чему реализуются механизмы пластичности мозга (Разумникова и др., 2009; Gazzaniga, 1995; Gratton et al., 1992; Krause et al., 2000). Параметры альфа-ритма меняются при изменении функциональных состояний мозга, при этом повышение его амплитуды свидетельствует об усилении тормозных процессов, а понижение — об увеличении активационных (Данилова, 2001; Костанов и др., 2012; Klimesh, 1999; Owen et al., 2006). При одновременной регистрации ЭЭГ и фМРТ обнаружено, что периодическое снижение мощности альфа-колебаний ЭЭГ в покое соответствует усилению общего уровня мозговой активности (Gazzaniga, 1995; Klimesh, 1999). По мнению О.М. Базановой (2009), именно альфа-активность организует биоэлектрическую активность целого мозга.

Известно, что наиболее высокая амплитуда ЭЭГ отмечена для альфа-активности, регистрируемой в теменно-затылочных отведениях, особенно в О1 и О2, причем в отведении О2 (как менее активированном) в фоне в большинстве случаев она выше, чем в О1 (Русалова, 2014). Показаны связи альфа-ритма с индивидуальными особенностями исследуемых лиц (Базанова, Афтанас, 2007; Разумникова и др., 2009; Русалова, Кислова, 2011), что свидетельствует о его важной роли в поведении человека, а именно в организации таких основополагающих процессов мозга, как внимание, память, мотивации и эмоции. Принято считать, что альфа-ритм играет ведущую роль в организации тормозного процесса вследствие влияния тонических таламических импульсов (Базанова, 2009).

Обнаруженная в настоящем исследовании у самоконтрольных лиц более высокая мощность альфа-осцилляций, а также более высокая когерентность дельта-ритма, который связывают с понижением активационных процессов, свидетельствует о возможности тормозного контроля их поведения. В то же время более высокая мощность бета2-ритма, который, как известно, является показателем увеличения активационных процессов, может свидетельствовать о том, что у импульсивных лиц преобладает мотивационная и активационная составляющие побуждения, недостаточно обеспеченные возможностью интеграции различных аспектов поступающей информации, что и создает, по видимому, особенности их темперамента, а именно быстрое принятие решений без соответствующего анализа их последствий.

Заключение

Таким образом, нами предложена новая методика процедур дискриминантного анализа с использованием Программного комплекса анализа и топографического картирования электрической активности мозга «Brainsys», которая дает исследователю мощное и гибкое средство для проведения дискриминантного анализа ЭЭГ в два класса, позволяющее в самые короткие сроки без рутинных процедур проверить множество вариантов дискриминантных функций, получить быстрые предварительные оценочные результаты процедурой пошагового включения предикторов и провести углубленный дискриминантный анализ данных.

Литература

- Айвазян, С. А., Бухштабер, В. М., Енюков, И. С., Мешалкин, Л. Д. (1989). *Прикладная статистика*. М.: Финансы и статистика.
- Андерсон, Т. (1963). *Введение в многомерный статистический анализ*. М.: Государственное издательство физико-математической литературы.
- Аффифи, А., Эйзен, С. (1982). *Статистический анализ*. М.: Мир.
- Базанова, О. М. (2009). Современная интерпретация альфа-активности электроэнцефалограммы. *Успехи физиологических наук*, 40(3), 32–53.
- Базанова, О. М., Афтанас, Л. И. (2007). Индивидуальные показатели альфа-активности электроэнцефалограммы и невербальная креативность. *Российский физиологический журнал имени И.М. Сеченова*, 93(1), 14–26.
- Бендат, Дж., Пирсол, А. (1974). *Измерение и анализ случайных процессов*. М.: Мир.
- Воронцов, К. В. (2004). Комбинаторный подход к оценке качества обучаемых алгоритмов. В кн. *Математические вопросы кибернетики* (вып. 13, с. 5–36). М.: Физматлит.
- Данилова, Н. Н. (2001). *Психофизиология*. М.: Аспект Пресс.
- Кирой, В. Н., Владимирский, Б. М., Асланян, Е. В., Бахтин, О. М., Миняева, Н. Р. (2010). Электрографические корреляты реальных и мысленных движений: спектральный анализ. *Журнал высшей нервной деятельности имени И.П. Павлова*, 60(5), 525–533.
- Костандов, Э. А., Черемушкин, Е. А., Ашкинази, М. Л. (2012). Динамика индуцированной корковой активности в паузах между целевым и пусковым стимулами при формировании установки на эмоциональное выражение лица. *Журнал высшей нервной деятельности имени И.П. Павлова*, 62(3), 311–321.
- Лапин, И. А., Митрофанов, А. А. (2017). Использование математического анализа ЭЭГ для дифференциальной диагностики биполярных и униполярных депрессий (на примере дискриминантного анализа показателей спектральной мощности, когерентности и межполушарной асимметрии). *Социальная и клиническая психиатрия*, 27(2), 70–73.
- Соловьева, Н. В., Кувшинова, Я. В., Чаусова, С. В., Вильянов, В. Б., Русалова М.Н., Олимпиева, С. П. (2019). Использование дискриминантного анализа ЭЭГ для диагностики шизофрении. *Журнал неврологии и психиатрии имени С.С. Корсакова*, 119(1), 44–50.
- Нюер, М. Р. (1992). Количественный анализ и топографическое картирование ЭЭГ: методики, проблемы, клиническое применение. *Успехи физиологических наук*, 23(1), 20–39.
- Разумникова, О. М., Тарасова, И. В., Вольф, Н. В. (2009). Особенности активации коры у лиц с высокой и низкой вербальной креативностью: анализ альфа 1,2-ритмов. *Журнал высшей нервной деятельности имени И.П. Павлова*, 59(5), 581–586.
- Рао, С. Р. (1973). *Линейные статистические методы и их применение*. М.: Наука.
- Русалова, М. Н. (2014). Асимметрия амплитуды альфа-ритма при мысленном воспроизведении эмоциональных образов. *Асимметрия*, 8(2), 24–40.
- Русалова, М. Н., Кислова, О. О. (2011). Электрофизиологические показатели распознавания эмоций в речи. *Успехи физиологических наук*, 42(2), 57–82.

Митрофанов Андрей Алексеевич — научный сотрудник, Федеральное государственное бюджетное научное учреждение Научный центр психического здоровья.

Сфера научных интересов: программирование, математико-статистическая обработка биосигналов.

Контакты: Brainsys@yandex.ru

Кичук Ирина Викторовна — доцент, кафедра неврологии, нейрохирургии и медицинской генетики, лечебный факультет, ГБОУ ВПО РНИМУ им. Н.И. Пирогова Минздрава России, кандидат медицинских наук.

Сфера научных интересов: нейрофизиология.

Контакты: mail2irina@mail.ru

Русалова Маргарита Николаевна — ведущий научный сотрудник, Федеральное государственное бюджетное научное учреждение Институт высшей нервной деятельности и нейрофизиологии РАН, доктор биологических наук.

Контакты: mgrusalova@ Rambler.ru

Чаусова Светлана Витальевна — заведующая кафедрой, кафедра общей патологии, МБФ ГБОУ ВПО РНИМУ им. Н.И. Пирогова Минздрава России, доктор медицинских наук, профессор.

Сфера научных интересов: патофизиология.

Контакты: svetlana_chau@mail.ru

Соловьева Надежда Валентиновна — генеральный директор, Акционерное общество «Научный центр персонализированной медицины», врач-психиатр.

Сфера научных интересов: психиатрия, нейропсихология.

Контакты: drsnv@yandex.ru

The Development of the Automated Discriminant Analysis of the EEG to Distinguish Two Classes

A.A. Mitrofanov^a, I.V. Kichuk^b, M.N. Rusalova^c, S.V. Chausova^b, N.V. Solov'eva^d

^a FSBSI Scientific center of mental health, 34, Kashirskoe shosse, Moscow, 115522, Russian Federation

^b SBEI HPE N.I. Pirogov RNRMU of Ministry of Health of Russian Federation, 1, Ostrovityanova, Moscow, 117997, Russian Federation

^c FSBSI Institute of higher nervous activity and neurophysiology RAS, 5a, Butlerova, Moscow, 117485, Russian Federation

^d Joint stock company «Scientific Center of Personalized Psychiatry», 20/3, Bolshaya Pochtovaya, Moscow, 105082, Russian Federation

Abstract

The authors propose a method of accelerated discriminant analysis procedures using the Software complex "Brainsys" to distinguish two classes. The software package allows us to consider up to 10 thousand spectral and other EEG parameters that can potentially act as predictors of a linear discriminant function (LDF). Provisionally, the EEG parameters can be transformed to an approximately normal distribution. Various procedures for selecting EEG parameters to satisfy the Fisher model are proposed, as well as procedures for searching for the most useful parameters — candidates for predictors. Procedures for stepwise inclusion of predictors and procedures for testing all possible combinations of predictors were considered. Special attention was paid to the problem of overfitting. Overestimates obtained from training samples can occur due to deviations from the Fisher model or multiple testing. The solution to the problem of overfitting should be found in the confirmation of assessment of the quality of discrimination in independent samples, at least in one. We also consider the issues of obtaining an estimate of the generalization performance (the ability to correctly predict on independent test samples) of a LDF. The method allows checking many variants of discriminant functions without routine calculations or get quick preliminary estimation results by applying the procedure of stepwise inclusion of predictors. A formula of LDF was obtained, which included EEG spectral parameters for differentiating persons with impulsive behavior from persons prone to self-control.

Keywords: impulsivity, self-control, electroencephalogram, discriminant analysis, generalization performance.

References

- Afifi, A. A., & Azen, S. P. (1982). *Statisticheskii analiz* [Statistical analysis]. Moscow: Mir. (in Russian; transl. of: Afifi, A. A., & Azen, S. P. (1979). *Statistical analysis. A computer oriented approach*. New York/San Francisco/London: Academic Press.)
- Aivazyan, S. A., Buchstaber, V. M., Yenyukov, I. S., & Meshalkin, L. D. (1989). *Prikladnaya statistika* [Applied statistics]. Moscow: Finansy i statistika. (in Russian)

- Anderson, T. (1963). *Vvedenie v mnogomernyi statisticheskii analiz* [An introduction to multivariate statistical analysis]. Moscow: Fizmatgiz. (in Russian; transl. of: Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York: John Wiley and Sons, Inc.)
- Bazanova, O. M. (2009). Electroencephalogram alpha activity: Modern interpretations. *Uspekhi Fiziologicheskikh Nauk*, 40(3), 32–53. (in Russian)
- Bazanova, O. M., & Aftanas, L. I. (2007). Individual alpha activity of electroencephalogram and non-verbal creativity. *Rossiiskii Fiziologicheskii Zhurnal imeni I.M. Sechenova*, 93(1), 14–26. (in Russian)
- Bendat, J. S., & Piersol, A. G. (1974). *Izmerenie i analiz sluchainykh protsessov* [Random data measurement and analysis procedures]. Moscow: Mir. (in Russian; transl. of: Bendat, J. S., & Piersol, A. G. (1971). *Random data; analysis and measurement procedures*. New York: Wiley-Interscience.)
- Daniilova, N. N. (2001). *Psikhofiziologiya* [Psychophysiology]. Moscow: Aspect-Press. (in Russian)
- Gazzaniga, M. S. (1995). *The cognitive neurosciences*. Cambridge: MIT Press.
- Gevins, A. S., & Remond, A. (Eds.). (1987). *Handbook of electroencephalography and clinical neurophysiology: Vol. 1: Methods of analysis of brain electrical and magnetic signals*. Amsterdam: Elsevier Science Publishers B.V.
- Gratton, G., Villa, A. E., & Fabiani, M. (1992). Functional correlates of a three-component spatial modes of the alpha-rhythm. *Brain Research Review*, 582, 159–173.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning. Data mining, inference, and prediction* (2nd ed.). New York: Springer.
- Kendall, M. G., & Stuart, A. (1968). *The advanced theory of statistics: Vol. 3. Design and analysis, and time-series*. London: Griffin.
- Kiroi, V. N., Vladimirkii, B. M., Aslanian, E. V., Bakhtin, O. M., & Miniaeva, N. R. (2010). Electrographic correlates of real and imaginary movements: spectral analysis. *Zhurnal Vysshei Nervnoi Deyatelnosti imeni I.P. Pavlova*, 60(5), 525–533. (in Russian)
- Klecka, W. R. (1980). *Discriminant analysis*. Newbury Park, CA: Sage Publications.
- Klimesh, W. (1999). EEG Alpha and Theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Review*, 29(2–3), 169–195.
- Kostandov, E. A., Cheremushkin, E. A., & Ashkinazi, M. L. (2012). Dynamics of induced EEG in the intervals between a target and trigger stimuli during formation of a set to emotionally negative facial expression. *Zhurnal Vysshei Nervnoi Deyatelnosti imeni I.P. Pavlova*, 62(3), 311–321. (in Russian)
- Krause, C. M., Sillanmaki, L., & Koivisto, M. (2000). The effects of memory load on event-related EEG desynchronization and synchronization. *Clinical Neurophysiology*, 111(11), 2071–2093.
- Lapin, I. A., & Mitrofanov, A. A. (2017). Use of mathematical analysis of EEG records for the purpose of differential diagnosis of uni- and bipolar depression (on example of discriminant analysis of power spectrum, coherence and hemispheric asymmetry parameters). *Sotsial'naya i Klinicheskaya Psikhatriya*, 27(2), 70–73. (in Russian)
- Ledermann, W., & Lloyd, E. (1984). *Handbook of applicable mathematics: Vol. VI: Statistics*. John Wiley & Sons Ltd.
- Mitrofanov, A. A., Kichuk, I. V., Solov'eva, N. V., Kuvshinova, Y. V., Chausova, S. V., Vileanov, V. B., Rusalova M.N., Olimpieva, S. P. (2019). The use of EEG discriminant analysis in the diagnosis of schizophrenia. *Zhurnal Nevrologii i Psihiatrii imeni S.S. Korsakova*, 119(1), 44–50. (in Russian)
- Nuwer, M. R. (1992). The quantitative analysis and topographic mapping of the EEG: the methods, problems and clinical use. *Uspekhi Fiziologicheskikh Nauk*, 23(1), 20–39. (in Russian)

- Owen, F. M., Coleman, M. R., & Boly, M. (2006). Detecting awareness in the vegetative state. *Science*, 313(5792), 1402–1421.
- Rao, C. R. (1973). *Lineinye statisticheskie metody i ikh primenenie* [Linear statistical inference and its applications]. Moscow: Nauka. (in Russian; transl. of: Rao, C. R. (1965). *Linear statistical inference and its applications*. New York: John Wiley & Sons.)
- Razumnikova, O. M., Tarasova, I. V., Wolf, N. V. (2009). Characteristics of cortical activity in persons with high and low verbal creativity: analysis of alpha_{1,2} rhythms. *Zhurnal Vysshei Nervnoi Deiatelnosti imeni I.P. Pavlova*, 59(5), 581–586. (in Russian)
- Rusalova, M. N. (2014). Asimmetriya amplitudy al'pha-ritma pri myslennom vosproizvedenii emotional'nykh obrazov [Asymmetry of the amplitude of alpha-rhythm at mental reproduction of emotional images]. *Asimmetriya*, 8(2), 24–40. (in Russian)
- Rusalova, M. N., & Kislova, O. O. (2011). Psychophysiological indexes of recognition of emotions in speech. *Uspekhi Fiziologicheskikh Nauk*, 42(2), 57–82. (in Russian)
- Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.
- Vorontsov, K. V. (2004). Kombinatornyi podkhod k otsenke kachestva obuchaemykh algoritmov [Combinatorial approach toward assessment of the quality of learning algorithms]. In *Matematicheskie voprosy kibernetiki* [Mathematical issues of cybernetics] (Iss. 13, pp. 5–36). Moscow: Fizmatlit. (in Russian)

Andrey A. Mitrofanov — Research Fellow, Federal State Budgetary Scientific Institution Mental Health Research Center.
Research Area: programming, mathematical and statistical processing of biosignals.
E-mail: Brainsys@yandex.ru

Irina V. Kichuk — Associate Professor, N.I. Pirogov Russian National Research Medical University, PhD in Medicine.
Research Area: neurophysiology.
E-mail: mail2irina@mail.ru

Margarita N. Rusalova — Lead Research Fellow, Federal State Budgetary Scientific Institution Institute of Higher Nervous Activity and Neurophysiology, Russian Academy of Sciences, DSc in Biology.
Research Area: neurophysiology.
E-mail: mrusalova@rambler.ru

Svetlana V. Chausova — Professor, N.I. Pirogov Russian National Research Medical University, DSc in Medicine.
Research Area: pathophysiology.
E-mail: svetlana_chau@mail.ru

Nadezhda V. Solov'eva — General Director, Scientific Center of Personalized Medicine, psychiatrist.
Research Area: psychiatry, neuropsychology.
E-mail: drsnv@yandex.ru