

подбору людей, однако окончательный выбор и ответственность всегда остаются за лицом, принимающим решения.

## Литература

1 Осипенко Н. Б. Пример «выращивания» регрессионной модели социального явления на базе критерия правдоподобности ее интерпретации / Н. Б. Осипенко, А. Н. Осипенко, К. А. Осипенко // Проблемы физики, математики и техники. – 2013. – № 4(17). – 85–88 с.

2 Осипенко, К. А. Метод регрессионного моделирования продолжительности жизни по дате рождения / К. А. Осипенко, Н. Б. Осипенко // Творчество молодых 2012: сборник научных работ студентов и аспирантов УО «ГГУ им. Ф. Скорины»: в 2 ч. / Гомельский гос. ун-т им. Ф.Скорины; отв. ред. О. М. Демиденко. – Гомель, 2012. – Ч. 1. – С. 194–197.

УДК 519.25

К. А. Осипенко

### ОБ ИСПОЛЬЗОВАНИИ АПРИОРНЫХ ЗНАНИЙ ПРИ РАСПОЗНАВАНИИ ГРУПП РИСКА СМЕРТНОСТИ НА ОСНОВЕ ПАСПОРТНЫХ ДАННЫХ

*В статье развивается предложенная в предыдущих работах идея массовой экспресс-диагностики групп риска смертности жителей в рамках трехэтапного скрининга здоровья населения Беларуси. В качестве доказательства работоспособности этой идеи было показано, что используя только паспортные данные уже для значительной части людей можно распознавать обобщенную группу риска. С практической точки зрения знание своей основной группы риска поможет человеку выстроить оптимальную стратегию здорового образа жизни.*

**Введение.** В статье [1] авторами был описан вариант трехэтапного скрининга здоровья населения Беларуси, где на первом этапе предлагается осуществить превентивную экспресс-диагностику групп риска смертности обследуемого человека с помощью стандартной информации из поликлинической базы данных. В качестве доказательства работоспособности этой идеи было показано, что используя только паспортные данные (дата рождения, фамилия, имя, отчество, семейное положение, количество детей), уже для значительной части людей можно распознавать обобщенную группу риска. В первую группу входят люди с высокой вероятностью смерти с быстрой потерей жизнеспособности организма (инфаркты, инсульты, несчастные случаи), во вторую – с медленной потерей жизнеспособности (онкологическая или иная хроническая болезнь). С практической точки зрения знание своей основной группы риска поможет человеку выстроить правильную стратегию профилактического поддержания своего здоровья. В связи с этим в настоящей статье развивается тема прогноза соответствующих групп риска для живущих людей.

**Описание исходных данных для статистического анализа.** В качестве исходного материала для пробного исследования возможностей использования паспортных сведений на первом этапе скрининга послужила выборка из 88 знаменитых людей преимущественно из артистической среды. По каждому человеку из интернета брались данные в виде строки: имя; фамилия; страна, в которой жил человек; тип

менталитета; день, месяц и год рождения; день, месяц и год смерти; количество жен (мужей); количество детей (включая приемных); основной род занятий по жизни (профессия); общая причина смерти (заболевание сердечнососудистой системы – инфаркт или инсульт – 1, онкозаболевание – 2, другие болезни – 3, несчастные случаи – 4, преднамеренное убийство – 5, иное – 6); детализация подсистем организма в структуре общей причины смерти.

Для обработки данных использовался пакет «Statistica», в частности, его средства подготовки новых переменных путём того или иного преобразования исходных признаков, а также программы анализа вариантов, построения гистограмм, матриц корреляций признаков, классификации данных и дискриминантного анализа. Кроме того, в рамках этого пакета была написана программа обучения распознаванию групп риска.

**Перевод даты рождения, имени и фамилии в качественные признаки.** Перевод даты рождения в набор числовых качественных признаков осуществлен с помощью общеизвестного алгоритма Пифагора и описан в работе [2]. В итоге из даты рождения получают следующие признаки: число (цифра) даты (получено путем поэтапного сложения всех ее цифр пока в сумме не получится одна цифра); количества встречаемости цифр 0, 1, ..., 9 в рабочих числах алгоритма Пифагора [2]; образованные из предыдущих десяти номинальных признаков бинарные признаки, показывающие наличие или отсутствие, например, ноль – двоек, одной – двойки и т.д. в квадрате Пифагора.

Для перевода имени и фамилии в признаки для распознавания групп риска нумерология предлагает множество различных соответствий букв и цифр. В настоящей работе использована числовая азбука [3], с помощью которой можно вместо буквы ставить числа: А-1, Б-2, В-3, Г-4, Д-5, Е-6, Ж-7, З-8, И-9, К-10, Л-20, М-30, Н-40, О-50, П-60, Р-70, С-80, Т-90, У-100, Ф-200, Х-300, Ц-400, Ч-500, Ш-600, Щ-700, Ю-800, Я-900, Э-1000. В работе [3] приведена таблица содержательных значений этих числовых кодов. Например, для известного баснописца Ивана Крылова кодирование имеет следующий вид: И-9, В-3, А-1, Н-40. В сумме 53; К-10, Р-70, Ы-0, Л-20, О-50, В-3. В сумме 153. Складываем обе полученные суммы ( $53 + 153 = 206$ ). Полученное число ищем из приведенной в [3] таблице. Но так как точно числа 206 в ней нет, то его следует разбить на  $200 + 6$ . Для числа 200 находим: хладнокровие, слабохарактерность, для 6 – труд, свободолюбие, успех. Как видно из биографии знаменитого баснописца, все эти качества были налицо. Теперь поступим с числом 206 иначе. Сложим его цифры до двузначного или однозначного вида и снова обратимся к таблице содержательной интерпретации чисел: 8 – величие, кротость, справедливость.

При наличии обучающей выборки в несколько сотен тысяч человек можно было бы воспользоваться полностью описанной здесь кодировкой. В нашем же пробном примере применялась только однозначная кодировка. При этом были образованы группы бинарных признаков для имени, для фамилии и для имени вместе с фамилией. Соответствующие числа (цифры) преобразовывались в последовательность из девяти бинарных признаков. Например, у семерки на седьмом месте в такой последовательности стоит единица, а на других местах – нули.

**Алгоритм распознавания групп риска.** В связи с небольшим объемом пробной выборки вместо четырех общих групп риска были выделены две. В первую группу вошли умершие от сердечнососудистых заболеваний и от несчастных случаев. Для нее характерна быстрая потеря жизнеспособности организма. Во вторую группу вошли те, у кого такая потеря шла постепенно – это умершие от онкозаболеваний и других хронических болезней. Также из-за недостаточного объема исходной выборки не удалось осуществить полноценное выделение ее однородных подвыборок, в частности, использовать для этого признаки пола, менталитета и продолжительности жизни (ранние смерти и смерти в зрелом возрасте). Сам алгоритм распознавания групп риска разбит на два этапа.

Этап 1. Обучение распознаванию групп риска.

1. Выбор информативных признаков.

1.1. Построение матрицы взаимных корреляций по всем признакам, включая целевой признак принадлежности к первой или второй обобщенным группам риска.

1.2. Выбор признаков, имеющих более или менее значимую связь с целевым свойством (в качестве критерия выбора информативного признака в нашем случае бралось условие:  $|r| \geq 0,1$ ). Положительная корреляция говорит о том, что признак в большей степени «голосует» за вторую группу риска и, наоборот, отрицательная корреляция говорит о предпочтении первой группы риска.

2. Формирование дискриминантной функции.

Пусть  $P_1 = \{p_1^1, p_2^1, \dots, p_{n_1}^1\}$  – множество признаков, голосующих за первую группу риска, а  $P_2 = \{p_1^2, p_2^2, \dots, p_{n_2}^2\}$  – множество признаков, голосующих за вторую группу риска.

Вероятность  $B_1$  отнесения объекта  $x$  к классу 1 вычисляется по формуле:

$$B_1 = (\sum_{i=1}^{n_1} x_i^1 / n_1) / ((\sum_{i=1}^{n_1} x_i^1) / n_1 + (\sum_{j=1}^{n_2} x_j^2) / n_2),$$

где  $x_i^1, (i = 1, \dots, n_1)$  – значения признаков из множества  $P_1$  для объекта  $x$ ;

$x_j^2, (j = 1, \dots, n_2)$  – значения признаков из множества  $P_2$  для объекта  $x$ .

Соответственно,  $B_2 = 1 - B_1$  – вероятность отнесения объекта  $x$  к классу 2.

Этап 2. Экзамен алгоритма распознавания групп риска.

Обычно экзамен проводится для объектов, не участвовавших при построении дискриминантной функции. Принятие решения об отнесении объекта к группе риска имеет вид:

$$R = \begin{cases} 1, & \text{если } B_1 - B_2 \geq \alpha; \\ 2, & \text{если } B_2 - B_1 \geq \alpha; \\ 0, & \text{если } |B_1 - B_2| < \alpha, \end{cases}$$

где  $0 \leq \alpha \leq 0.5$  (в нашем примере  $\alpha = 0.2$ ).

Если  $R = 0$ , то программа не может различить группу риска для объекта  $x$ .

**Результаты распознавания групп риска по исходной выборке.** Оценка качества дискриминантной функции проводилась по исходной выборке. При этом ошибка первого рода (отнесение объекта первого класса ко второму) составила 10/55, ошибка второго рода – 1/33. Доля отказов – 39/88.

Высокую ошибку распознавания первой группы (в основном это смерти от сердечнососудистой болезни) можно объяснить неполнотой диагноза смертельного исхода, особенно в зрелом возрасте, когда инфаркт или инсульт становится следствием какой-либо хронической болезни. Часто это не афишируется даже для знаменитостей и никак не отражается в диагнозах смерти.

**Результаты прогноза групп риска.** С целью проверки практической пригодности предлагаемого метода был осуществлен прогноз групп риска для выборки из ныне здравствующих публичных личностей из артистической среды (Таблица 1).

Таблица 1 – Вероятности отнесения к первой и второй группам риска

Фамилия, имя		I группа	II Группа
Питт	Уильям	0,42	0,58
Джоли	Анжелина	0,18	0,82
Сомерхолдер	Иэн	0,78	0,22
Касас	Марио	0,42	0,58
Вальверде	Мария	0,35	0,65
Беллуччи	Моника	0,42	0,58
Добрев	Нина	0,64	0,36
Диас	Камерон	0,42	0,58
Турман	Ума	0,31	0,69
Уотсон	Эмма	0,47	0,53
Ди Каприо	Леонардо	0,47	0,53
Волочкова	Анастасия	0,35	0,65
Ротару	София	0,78	0,22
Пугачёва	Алла	0,47	0,53
Киркоров	Филипп	0,26	0,74
Андерсон	Памела	0,28	0,72
Пэлтроу	Гвинет	0,73	0,27
Энистон	Дженнифер	0,64	0,36
Клуни	Джордж	0,54	0,46
Варум	Анжелика	0,47	0,53
Белан	Виктор	0,47	0,53
Брежнева	Вера	0,47	0,53
Седокова	Анна	0,26	0,74

Исходя из этой таблицы вероятностей, можно делать предварительные выводы о генетической предрасположенности организма человека к тому или иному риску смерти. Так, ко второй группе риска (смерть с медленной потерей жизнеспособности, в частности, в связи с онкозаболеванием) можно отнести Анжелину Джоли, Марию Вальверде, Уму Турман, Анастасию Волочкову, Филиппа Киркорова, Памелу Андерсон и Анну Седокову. Соответственно, этим людям желательно порекомендовать, с одной стороны, аккуратнее относиться ко всему, что негативно воздействует на иммунную и гормональную системы организма, и с другой – подобрать с помощью специалистов оптимальную технологию здорового образа жизни, включающую соответствующую диету и физические упражнения, а также периодическую корректировку вредных ментально-психологических убеждений и ощущений, ответственных за инициацию и развитие онкозаболевания. К первой группе риска (смерть с быстрой потерей жизнеспособности, в частности, в связи с проблемами сердечнососудистой системы) можно отнести Иэн Сомерхолдер, Нину Добрев, Софию Ротару, Гвинет Пэлтроу и Дженифер Энистон. Для этих людей потребуются аналогичные рекомендации, но уже с учетом специфики формирования проблем в работе сердечнососудистой системы. По остальным личностям данной выборки артистов в рамках данного исследования решение о принадлежности к той или иной группе не принимается. Это может быть связано либо с недостаточной информативностью исходных данных, либо с наличием третьего варианта генетической предрасположенности – с более или менее равновероятным отнесением к описанным двум группам риска. Известно также [4], что в течение жизни, в результате активного медикаментозного лечения, человек может перейти из первой группы риска (с низким гуморальным иммунитетом) во вторую (с низким клеточным иммунитетом) и наоборот. В связи с этим для более качественного прогноза групп риска целесообразно исходную обучающую

выборку подбирать так, чтобы у них в течение жизни не происходила замена одного типа ослабления иммунитета другим.

### Литература

- 1 Осипенко, К. А. О возможности использования паспортных данных в статистическом анализе причин смертности населения / К. А. Осипенко, Н. Б. Осипенко // Творчество молодых 2013: сборник научных работ студентов и аспирантов УО «ГГУ им. Ф. Скорины»: в 3 ч. / Гомельский гос. ун-т им. Ф. Скорины; отв. ред. О. М. Демиденко. – Гомель, 2013. – Ч. 1. – С. 163–167.
- 2 Осипенко, К. А. Метод регрессионного моделирования продолжительности жизни по дате рождения / К. А. Осипенко, Н. Б. Осипенко // Творчество молодых 2012: сборник научных работ студентов и аспирантов УО «ГГУ им. Ф. Скорины»: в 2 ч. / Гомельский гос. ун-т им. Ф. Скорины; отв. ред. О. М. Демиденко. – Гомель, 2012. – Ч. 1. – С. 194–197.
- 3 Хигир, Б. Ю. Число имени / Б. Ю. Хигир. – СПб.: Астрель, 2008. – 42 с.
- 4 Лободин, В. Т. Формула здоровья / В. Т. Лободин. – СПб.: ИД Невский проспект, 1999. – 187 с.

УДК 539.2

*А. И. Остапенко*

### ЧИСЛЕННОЕ ИССЛЕДОВАНИЕ СТРУКТУР, ОБРАЗОВАННЫХ ФУЛЛЕРЕНАМИ

*Данная работа посвящена изучению углеродных наноструктур, образованных молекулами эндофуллеренов, моделированию таких наноструктур в программе NureChem, расчету оптимальных геометрий выбранных конформаций, энергии образования и оценке возможности существования таких структур.*

Одними из самых известных исследуемых объектов нанофизики являются фуллерены. Они представляют собой систему связанных атомов углерода, образующих сферический каркас. Самым распространенным, изученным и стабильным является фуллерен  $C_{60}$ , форма которого состоит из 20 шестиугольников и 12 пятиугольников. Существуют и другие фуллерены с большим и меньшим количеством атомов. Конформации с числом атомов от 20 до 60 называются низшими фуллеренами, остальные (до 540) – высшими (рисунок 1). Для оценки масштаба изображения на рисунке приведена горизонтальная линия, длина которой соответствует расстоянию  $65 \text{ \AA}$ .

Наряду с другими, для изучения таких нанообъектов используют методы компьютерного моделирования, разработанные с учетом экспериментальной информации и различных методов численного решения многочастичного уравнения Шредингера. Одной из самых известных программ, реализующих различные методы моделирования и расчета структуры нанообъектов, является NureChem. Для каждого метода в программе имеется набор различных параметризаций, позволяющих варьировать выбор варианта постановки задачи и набора базисных функций в зависимости от цели исследования и возможностей используемой вычислительной техники [1].