УДК 57.087 + 004.89 + 007.51

ОЦЕНКА ДЛИНЫ ОБУЧАЮЩЕЙ ПОСЛЕДОВАТЕЛЬНОСТИ В ЗАДАЧЕ РАСПОЗНАВАНИЯ ОБРАЗОВ (БИОИНДИКАЦИЯ)

Г.С. Розенберг

Институт экологии Волжского бассейна РАН, Тольятти, Россия
genarozenberg@yandex.ru

Аннотация

Задача создания системы распознавания образов распадается на ряд подзадач: формализации предметной области, формирования обучающей выборки, обучения системы распознавания, снижения размерности пространства признаков, собственно задача распознавания (по степени сходства распознаваемого объекта с обучающей выборкой), контроля качества распознавания, адаптации, обратной задачи распознавания, кластерного и конструктивного анализа, когнитивного анализа. В статье рассматривается формализация одной из подзадач (формирования обучающей выборки). С помощью предложенной вероятностной модели сделан вывод о «почти линейной зависимости» длины обучающей последовательности и размерности пространства признаков. Получена оценка длины обучающей последовательности для реалистичных значений параметров модели.

Ключевые слова: *биоиндикация, распознавание образов, длина обучающей последовательности, случайная величина, решающее правило.*

Цитирование: *Розенберг, Г.С.* Оценка длины обучающей последовательности в задаче распознавания образов (биоиндикация) / Г.С. Розенберг // Онтология проектирования. – 2017. – Т. 7, №2(24). - С. 207-215. – DOI: 10.18287/2223-9537-2017-7-2-207-215.

Введение

Содержательная (а не формальная) постановка задачи распознавания образов появилась в конце 50-х годов прошлого века и заключалась в том, чтобы построить систему, способную обучаться классификации ситуаций так же, как это делают живые существа. Такое широкое понимание проблемы привело к возникновению различных направлений этих исследований [1, 2]: одни считали главным построение модели восприятия [3], другие видели проблему в том, чтобы используя априорные сведения о свойствах образов, найти такое их описание, при котором отыскание принципа классификации не составляет труда [4], третьи понимали задачу распознавания образов как задачу минимизации риска в специальном классе решающих правил [5], четвёртые определяли распознавание образов как процесс, не требующий точного описания и основанный на обнаружении сходства объектов [6-8], а в работе [9] в деятельности по распознаванию образов находят сходство с процессом проектирования объектов. Эта «пестрота» подходов к задаче распознавания образов объясняется тем, что разные авторы различно представляют как само понятие «образ», так и процесс распознавания. Главным препятствием, стоящим на пути исследователей в этой обширной области, является отсутствие чёткого понимания того, какие процессы происходят при обучении человека.

Распознавание представляет собой информационный процесс, реализуемый некоторым преобразователем информации (системой распознавания), имеющим вход и выход. На вход системы подаётся информация о том, какими признаками обладают предъявляемые объекты; на выходе системы отображается информация о том, к каким классам (обобщённым образам)

отнесены распознаваемые объекты. При создании и эксплуатации автоматизированной системы распознавания образов решается ряд задач.

Рассмотрим кратко основные из них [10]:

1) задача формализации предметной области (задача кодирования);
2) задача формирования обучающей выборки (база данных, содержащая описание объектов в пространстве признаков, дополненная информацией о принадлежности этих объектов к определенным классам распознавания);
3) задача обучения системы распознавания (обучающая выборка используется для формирования обобщенных образов классов распознавания);
4) задача снижения размерности пространства признаков (после обучения системы распознавания можно определить для каждого признака его ценность для решения задачи распознавания; наименее ценные признаки могут быть удалены из системы признаков; затем система распознавания должна быть обучена заново и этот процесс может повторяться, т.е. быть итерационным);
5) собственно задача распознавания (по степени сходства распознаваемого объекта с обучающей выборкой);
6) задача контроля качества распознавания (адекватность; определение фактической средней вероятности ошибки по всем классам распознавания, а также вероятности ошибки при отнесении распознаваемого объекта к определённому классу; результаты распознавания должны интерпретироваться с учётом имеющейся информации о качестве распознавания);
7) задача адаптации (если контроль качества распознавания неудовлетворителен, то необходимо переформатировать распознаваемую и обучающую выборки, вновь решить задачу обучения системы распознавания, стремясь повысить адекватность классификации распознаваемых объектов);
8) обратная задача распознавания (для данного класса распознавания системой устанавливается, какие признаки наиболее характерны для объектов данного класса);
9) задачи кластерного и конструктивного анализа (результатом кластерного анализа является классификация объектов по кластерам);
10) задача когнитивного анализа (анализ пространства признаков по сходству классов распознавания: наличие одного признака у разных классов распознавания или различных признаков у разных классов вносит определённый [интерпретируемый] вклад в их сходство и различие).

В данной статье более подробно формализуется и решается задача 2.

1 Оценка длины обучающей последовательности

Задача распознавания образов «с учителем» может быть сформулирована в следующем виде: пусть для l объектов $\{o_i\}$ ($i=\overline{1,l}$; назовем их *обучающей последовательностью*) известна принадлежность каждого объекта к одному из классов $\{A_j\}$, где $j=\overline{1,k}$. Каждый объект задается набором N признаков и, следовательно, может быть представлен точкой в N-мерном пространстве признаков X. В пространстве X строится разделяющая функция (решающее правило) Φ таким образом, чтобы она разбила X на k непересекающихся подпространств $\{X_j\}$, в каждом из которых будут находиться объекты $\{o_i\}$ только одного класса; например, для $k=2$ имеем: $\Phi(o_i)\geq 0$ для $o_i \in A_1$ и $\Phi(o_i)<0$ для $o_i \in A_2$, где $A_1 \cup A_2$, $A_1 \cap A_2 = \varnothing$. Следует ожидать, что качество решающего правила, представляющее собой степень соответствия построенного разбиения пространства X на подпространства $\{X_j\}$ действительным классам $\{A_j\}$,

должно быть функцией длины обучающей последовательности l, размерности пространства признаков N и числа классов обучения k, то есть $\Phi = \Phi(l, N, k, o_i)$.

Существует мнение, что в качестве исходных признаков необходимо задавать всё, что только можно заподозрить в информативности. С другой стороны, замечено, что при практическом решении задач распознавания образов увеличение размерности пространства признаков не только не улучшает качество распознавания, но, в некоторых случаях, ухудшает его [11, 12]. Отсюда возникает задача оценки длины обучающей последовательности в зависимости от размерности пространства признаков N. Приводимое в литературе решение этой задачи [13, 14] содержит лишь грубую оценку, что не позволяет воспользоваться полученными результатами во многих теоретических и прикладных работах. Данная статья ставит своей целью получение более точных формул решения этой задачи.

Пусть Z – некоторая случайная величина, элементарными событиями которой являются $\{Z_i\}$ с вероятностями $\{p_i\}$, и пусть S – некоторая совокупность элементарных событий из множества $\{Z_i\}$. Пусть предстоит провести l независимых повторных наблюдений над случайной величиной Z. Через $\nu_A^{(l)} = n_A / l$ обозначим частоту появления события $A \in S$ для некоторой выборки длиной l, где n_A – число элементов выборки, принадлежащих A. Классическая теорема Бернулли утверждает, что при фиксированном событии A последовательность $\nu_A^{(l)}$ подчиняется биномиальному закону распределения и сходится к $p(A)$ по вероятности при $l \to \infty$, то есть для любого $\varepsilon > 0$ справедливо

$$P\left\{\left|\nu_A^{(l)} - P(A)\right| > \varepsilon\right\} \xrightarrow[l\to\infty]{} 0 .$$

Обозначим через $Y(l) = \sup\limits_{A\in S}\left|\nu_A^{(l)} - P(A)\right|$ случайную величину, представляющую собой наибольшее уклонение частоты $\nu_A^{(l)}$ от вероятности события A по всем $A \in S$. Если для любого $\varepsilon > 0$ справедливо

$$P\{Y(l) > \varepsilon\} \xrightarrow[l\to\infty]{} 0 ,$$

то можно говорить о равномерной сходимости частоты к своей вероятности [15-17]. Событие $Y(l) > \varepsilon$ произойдет в том случае, если хотя бы для одного из событий $A_j \in S$, где $j = \overline{1,T}$ (T – число наблюдений за реализацией Z в классе A_j), будет справедливо $\left|\nu_A^{(l)} - P(A)\right| > \varepsilon$. Так как наблюдения над Z проводятся независимо друг от друга, используя теорему о сложении вероятностей [15, 17], получим:

$$P\{Y(l) > \varepsilon\} \le \sum_{j=1}^{T} P\left\{\left|\nu_A^{(l)} - P(A_j)\right| > \varepsilon\right\} \le T \cdot P\left\{\left|\nu_A^{(l)} - P(A)\right| > \varepsilon\right\}, \quad (1)$$

где в качестве A взято событие, для которого $P\left\{\left|\nu_A^{(l)} - P(A_j)\right| > \varepsilon\right\}$ принимает наибольшее значение. Учитывая, что $\nu_A^{(l)}$ подчиняется биномиальному закону распределения, при больших l справедлива следующая оценка [6]:

$$P\left\{\left|\nu_A^{(l)} - P(A)\right| > \varepsilon\right\} \le \sqrt{\frac{2P(A)\left[1-P(A)\right]}{\pi\, l\, \varepsilon^2}}\, \exp\left(-\frac{l\,\varepsilon^2}{2P(A)\left[1-P(A)\right]}\right).$$

Подставляя эту оценку в (1) и полагая $P(A) = P$, получим:

$$P\{Y(l) > \varepsilon\} \le T \cdot \sqrt{\frac{2P(1-P)}{\pi l \varepsilon^2}} \exp\left(-\frac{l\varepsilon^2}{2P(1-P)}\right).$$

Если потребовать, чтобы вероятность $P\{Y(l) > \varepsilon\}$ не превосходила некоторого достаточно малого δ, то из уравнения

$$T \cdot \sqrt{\frac{2P(1-P)}{\pi l \varepsilon^2}} \exp\left(-\frac{l\varepsilon^2}{2P(1-P)}\right) = \delta \quad (2)$$

при заданных P и ε можно определить l.

Проведённые выше рассуждения справедливы для фиксированного конечного T. Если представить T как некоторую функцию длины выборки l и размерности пространства элементарных событий N, то есть $T = T(l, N)$, то условие равномерной сходимости (1) будет выполнено только в том случае, если $T = T(l, N)$ будет возрастать с ростом l медленнее, чем $\exp\left(\frac{l\varepsilon^2}{2P(1-P)}\right)$.

В общем случае, если не учитывать вид решающего правила, число всевозможных классификаций l объектов на k классов будет k^l, так как каждый объект может принадлежать любому из k классов. Однако в зависимости от положения объектов в пространстве X и от вида решающего правила Φ, число классификаций может быть значительно меньше. В ряде работ [18, 19] получена точная формула числа линейных дихотомий, то есть числа разбиения совокупности объектов на два класса с помощью гиперплоскостей:

$$T_2(l,N) = \begin{cases} 2^l, & \text{для } N \ge l, \\ 2 \cdot \sum_{i=0}^{N} C_{l-1}^i, & \text{для } N < l. \end{cases}$$

Таким образом, подставляя в (2) значение $T = T_2(l, N)$ для $l > N$, получим

$$\sum_{i=1}^{N} C_{l-1}^i = \frac{\delta\varepsilon}{2}\sqrt{\frac{\pi l}{2P(1-P)}} \exp\left(\frac{l\varepsilon^2}{2P(1-P)}\right). \quad (3)$$

Следовательно, с вероятностью $1 - \delta$ можно утверждать, что частота правильной классификации N-мерных объектов с помощью гиперплоскости будет отличаться от вероятности P не более, чем на $\varepsilon > 0$, если длина обучающей последовательности $l > l_0$, где l_0 является решением уравнения (3). Анализ этой связи позволяет сделать вывод о «почти линейной зависимости» длины обучающей последовательности и размерности пространства признаков. Например, для $\delta = 0{,}2$, $P = 0{,}95$, $\varepsilon = 0{,}2$ и $N < 250$ можно считать $l \approx 6N$.

В работе [20] показано, что необходимость в разделении объектов кусочно-линейными поверхностями возникает лишь в самых сложных случаях взаимного расположения этих объектов и, следовательно, полученная оценка является завышенной. Это подтверждает результаты вычислительных экспериментов по редукции длины обучающей последовательности, представленные в [21, 22]. Таким образом, в зависимости от степени «компактности» распознаваемых объектов в пространстве признаков в практической работе можно пользоваться числом l из интервала $3N \le l \le 6N$ (для указанных выше δ, P и ε).

2 Экологическая интерпретация (биоиндикация)

Содержание принципа антропоцентризма исторически менялось, исходя из понимания сущности человека (в рамках гуманитарных представлений различных философских школ и учений). Однако он всегда находился «во главе угла» в отношении истории развития нормирования (наложение граничных условий [нормативов] как на само воздействие, так и на факторы среды, отражающие и воздействие, и отклики экосистем): значительно ранее прочих были установлены нормативы приемлемых именно для человека условий среды (прежде всего, производственной). Тем самым было положено начало работам в области санитарно-гигиенического нормирования. Однако человек – не самый чувствительный из биологических видов и принцип «защищён человек – защищены и экосистемы», вообще говоря, неверен. Таким образом, экологическое нормирование является ключевой проблемой в формировании экологической безопасности.

В принципе, можно изучать фактор среды, проводя его прямые измерения-анализы. Однако, если для некоторых факторов эти измерения сравнительно просты (например, высота над уровнем моря или интенсивность γ-излучения), то для других представляют значительную сложность (например, засолённость почвы) или просто невозможны при однократном наблюдении (увлажнение почвы или изменяющаяся со временем концентрация некоторого загрязнителя в водной среде). Кроме того, как бы мы ни снижали уровень отрицательного воздействия, например, на водные массы, инструментальными методами невозможно контролировать присутствие *всех* загрязнителей. Занятие это очень трудоёмкое и финансово крайне затратное. Нужен постоянный контроль качества водной среды, а его может обеспечить только гидробиологический мониторинг. С другой стороны, связь между живыми организмами и факторами среды носит случайный характер и для её изучения необходимо привлекать вероятностное моделирование и строгие статистические методы.

Экологическое нормирование не является подменой санитарно-гигиеническому нормированию, а, в определённом смысле, дополняет его, ужесточая применяемые стандарты. Например, экологическая индикация может дать сведения о степени и характере загрязнения, распределении загрязнения в водоёме, возможном состоянии водной экосистемы в сезонном масштабе. Из этого следует, что вода, качество которой согласно экологическому контролю признано неудовлетворительным, вряд ли может использоваться для питьевых или хозяйственных целей, но экологически доброкачественная вода не всегда может быть признана пригодной с точки зрения здравоохранения. В последнем случае необходимы специфические микробиологические, токсикологические и химические тесты.

Одной из основ экологического нормирования является задача распознавания факторов среды по видам или группам видов-индикаторов (например, по растительности – геоботаническая индикация, фитоиндикация), актуальность которой не вызывает сомнений. Правда, почти полвека тому назад, крупнейшие фитоценологи отмечали наметившийся «застой» в количественных методах: «Как ни странно, но задачи фитоиндикации, вероятностные по своей природе, до сих пор решаются в основном без использования каких-либо статистических методов» [23, с. 141] и «Репрезентативные выборки, где независимо учитываются объект индикации и индикаторы, и строго статистические методы обработки данных – единственно возможный путь дальнейшего прогресса фитоиндиакационных исследований» [24, с. 1342]. С сожалением приходится констатировать, что ситуация за это время в биоиндикационных исследованиях кардинально не поменялась [25].

Среди причин объективного характера сложившейся ситуации можно назвать тот факт, что экосистемам, чаще всего, свойственен континуум по факторам среды и, следовательно, деление фактора на градации – процесс условный. В какой-то степени, нам «нельзя говорить о наилучшем (или оптимальном) решении, что, наоборот, характерно для задач в замкнутой

форме… При решении задач в открытой форме используются не только (и не столько) верифицируемые знания, главное для которых – их доказуемая истинность, но и аксиологические ценностно ориентированные знания» [26, с. 13]. Это обосновывает применение онтологического подхода [27, 28] в задачах распознавания образов, а проведённая выше оптимизация длины обучающей последовательности позволяет «объективизировать» процесс биоиндикации (в частности, геоботанической индикации факторов среды по растительности [29, 30]).

Заключение

Следует признать, что как и любая сложная система (сложный процесс), в которой участвует Человек [31], процесс биоиндикации (распознавания образов) не может быть полностью формализован без учёта онтологических принципов. Включение Человека со всем багажом его знаний в систему распознавания образов «на равных» (замена субъект-объектных отношений на субъект-субъектные [26]) способно придать задаче биоиндикации бо́льшую концептуализацию предметной области. «Однако "знать", это ещё не значит "делать". Поэтому приходится придерживаться ещё и ***этического рационализма***, согласно которому в основе поведения людей лежит (или должно лежать) рациональное начало; соответственно, знание о том, как необходимо поступать, является в данном случае достаточным условием нормативного поведения (*выделено автором – Г.Р.*)» [26, с. 15]. В качестве «рационального начала» в этой статье и выступает оценка длины обучающей последовательности в процедуре распознавания образов.

Благодарности

Автор благодарен профессорам Б.С. Флейшману, В.А. Виттиху и С.В. Смирнову за обсуждение некоторых проблем системологии. Работа выполнена при частичной финансовой поддержке Российского гуманитарного научного фонда (грант 16-13-63004-Самара) и Российского фонда фундаментальных исследований (грант 17-44-630113).

Список источников

[1] ***Сойфер, В.А.*** Методы компьютерной обработки изображений / В.А. Сойфер (ред.). – М.: Физматлит, 2001. – 784 с.

[2] ***Донской, В.И.*** Алгоритмические модели обучения классификации: обоснование, сравнение, выбор / Донской В.И. – Симферополь: ДИАЙПИ, 2014. – 228 с.

[3] ***Сочивко, В.П.*** Электронные опознающие устройства / Сочивко В.П. – М.: Энергия, 1964. – 56 с. (Сер.: Библиотека по автоматике. Вып. 91).

[4] ***Трапезников, В.А.*** Кибернетика и автоматическое управление / В.А. Трапезников // Вестн. АН СССР. – 1962. – № 5. – С. 33-42.

[5] ***Вапник, В.Н.*** Теория распознавания образов (теоретические проблемы обучения) / В.Н. Вапник, А.Я. Червоненкис. – М.: Наука, 1974. – 416 с.

[6] ***Бонгард, М.М.*** Моделирование процесса узнавания на цифровой счетной машине / М.М. Бонгард // Биофизика. – 1961. – Т. 4, № 2. – С. 17-23.

[7] ***Браверман, Э.М.*** О методе потенциальных функций / Э.М. Браверман // Автоматика и телемеханика. – 1965. – Т. 26, № 12. – С. 2205–2213.

[8] ***Айзерман, М.А.*** Теоретические основы метода потенциальных функций в задаче об обучении автоматов разделению входных ситуаций на классы / М.А. Айзерман, Э.М. Браверман, Л.И. Розоноэр // Автоматика и телемеханика. – 1964. – Т. 25, № 6. – С. 917-936.

[9] ***Боргест, Н.М.*** Распознавание образов при создании артефактов как метафора и как прикладные технологии онтологии проектирования / Н.М. Боргест // Онтология проектирования. - 2015. – Т.5, №1(15). – С. 19-29.

[10] ***Симанков, В.С.*** Адаптивное управление сложными системами на основе теории распознавания образов / В.С. Симанков, Е.В. Луценко – Краснодар: Кубан. гос. технол. ун-т, 1999. – 318 с.
[11] ***Харкевич, А.А.*** О выборе признаков при машинном опознании / А.А. Харкевич // Изв. АН СССР, сер. техн. киберн. – 1963. – № 2. – С. 3-9.
[12] ***Kanal, L.N.*** On dimensionality and sample size in statistical pattern classification / L.N. Kanal, B. Chandrasekaran // Pattern Recognition. – 1971. – V. 3, No. 3. – P. 225-234.
[13] ***Вапник, В.Н.*** Об одном классе алгоритмов обучения распознаванию образов / В.Н. Вапник, А.Я. Червоненкис // Автоматика и телемеханика – 1964. – Т. 25, № 6. – С. 937-945.
[14] ***Бабу, К.Ч.*** О применении потенциальной функции Башкирова, Бравермана и Мучника для выделения информативных признаков при распознавании образов / К.Ч. Бабу , С.Н. Калра // Автоматика и телемеханика. – 1972. – Т. 33, № 12. – С. 105-107.
[15] ***Смирнов, Н.В.*** О приближении плотностей распределения случайных величин / Н.В. Смирнов // Уч. зап. МГПИ им. В.П. Потемкина. – 1951. – Т. 16, вып. III. – С. 69-96.
[16] ***Слуцкий, Е.Е.*** Избранные труды: Теория вероятностей. Математическая статистика / Е.Е. Слуцкий – М.: Изд-во АН СССР, 1960. – 291 с.
[17] ***Вапник, В.Н.*** О равномерной сходимости частот появления событий к их вероятностям / В.Н. Вапник, А.Я. Червоненкис // Докл. АН СССР (ДАН). – 1968. – Т. 181, вып. 4. – С. 781-784.
[18] ***Нильсон, Н.*** Обучающие машины / Н. Нильсон. – М.: Мир, 1967. – 180 с.
[19] ***Флейшман, Б.С.*** Элементы теории потенциальной эффективности сложных систем / Б.С. Флейшман. – М.: Сов. радио, 1971. – 224 с.
[20] ***Лецкий, Э.К.*** Оптимизация решений при проектировании сетей сбора данных / Э.К. Лецкий // Вестн. МИИТ. – 1998. – Вып. 1. – С. 125-130.
[21] ***Розенберг, Г.С.*** Опыт приложения теории распознавания образов для оценки засоления почв по растительности / Г.С. Розенберг, Б.М. Миркин , С.Ю. Рудерман // Экология. – 1972. – № 6. – С. 31-34.
[22] ***Розенберг, Г.С.*** Редукция числа признаков и эффективность оценки почв по растительности при использовании методов распознавания образов / Г.С. Розенберг // Количественные методы анализа растительности: Материалы IV Всесоюз. совещания по проблеме «Применение количественных методов в анализе структуры растительности». – Уфа: БФАН СССР, 1974. – С. 36-39.
[23] ***Василевич, В.И.*** Второе совещание «Применение количественных методов при изучении структуры растительности» (Тарту, 1969) / В.И. Василевич // Ботан. журн. – 1970. – Т. 55, № 1. – С. 140-142.
[24] ***Миркин, Б.М.*** Рецензия «Теоретические вопросы фитоиндикации» / Б.М. Миркин // Ботан. журн. – 1972. – Т. 57, № 12. – С. 1342-1344.
[25] ***Шитиков, В.К.*** Количественная гидроэкология: методы, критерии, решения: в 2-х кн. / В.К. Шитиков, Г.С. Розенберг, Т.Д. Зинченко. – М.: Наука, 2005. – Кн. 1. 281 с.; Кн. 2. 337 с.
[26] ***Виттих, В.А.*** Парадигма ограниченной рациональности принятия решений: препринт / В.А. Виттих – Самара: ИПУСС РАН, 2009. – 26 с.
[27] ***Смирнов, С.В.*** Онтологический анализ предметных областей моделирования / С.В. Смирнов // Известия Самарского НЦ РАН. – 2001. – Т. 3, № 1. – С. 62-70.
[28] ***Смирнов, С.В.*** Онтологическое моделирование в ситуационном управлении / С.В. Смирнов // Онтология проектирования. – 2012. – № 2 (4). – С. 16-24.
[29] ***Розенберг, Г.С.*** Модели в фитоценологии / Г.С. Розенберг. – М.: Наука, 1984. – 240 с.
[30] ***Розенберг, Г.С.*** Введение в теоретическую экологию: в 2-х т.; изд. 2-е, исправ. и дополненное / Г.С. Розенберг. – Тольятти: Кассандра, 2013. – Т. 1. 565 с.; Т. 2. 445 с.
[31] ***Розенберг, Г.С.*** О простых, сложных и суперсложных системах / Розенберг Г.С. // Четверта міжнародна науково-практична конференція «Відкриті еволюціо-нуючі системи» (20-21 травня 2016 р.). Збірник праць: Частина 1. – Ніжин: ВНЗ ВП НУБіП України НАІ, 2016. – С. 228-233.

THE ESTIMATE OF THE LENGTH OF THE TRAINING SEQUENCE IN THE TASK OF PATTERN RECOGNITION (BIOINDICATION)

G.S. Rozenberg

Institute of Ecology of the Volga River Basin of the RAS, Togliatti, Russia
genarozenberg@yandex.ru

Abstract

Development of a pattern recognition system is comprised from a number of subtasks: formalization of the subject area, formation of a training sample, training of the recognition system, reduction of the dimensionality of the feature space, the problem of pattern recognition itself (based on the degree of similarity of the detected object with the training sample), recognition quality control, adaptation, inverse problem recognition, cluster and structural analysis, cognitive analysis. The article considers the formalization one of the subtasks (formation of the training sample). The conclusion about the "almost linear dependence" of the length of the training sequences and the dimensionality of the feature space is made using the proposed probabilistic model. Evaluation of the length of the training sequences for realistic values of model parameters is performed.

Key words: *bioindication, pattern recognition, the length of training sequences, a random variable, the decision rule.*

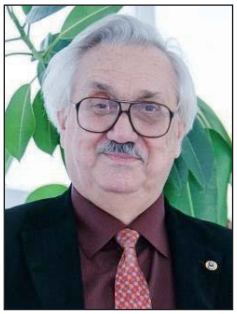
Citation: *Rozenberg GS.* Estimation of the length of the training sequence in the problem of pattern recognition (bioindication). Ontology of designing. 2017; 7(2): 207-215. DOI: 10.18287/2223-9537-2017-7-2-207-215.

References

[1] ***Soifer VA.*** Methods of Computer Image Processing [In Russian]. – Moscow: Fizmatlit, 2001. – 784 p.
[2] ***Donskoy VI.*** Algorithmic models of learning classification: rationale, comparison, selection [In Russian]. – Simferopol: DIP, 2014. – 228 p.
[3] ***Sochivko VP.*** Electronic Identifying Device [In Russian]. – Moscow: Energiya, 1964. – 56 p.
[4] ***Trapeznikov VA.*** Cybernetics and automatic control [In Russian]. Vestn. Acad. of Sci. of the USSR. – 1962. – No. 5. – P. 33-42.
[5] ***Vapnik VN., Chervonenkis AYa.*** Theory of Pattern Recognition (Theoretical Problems of Studying) [In Russian]. – Moscow: Nauka, 1974. – 416 p
[6] ***Bongard MM.*** Modeling of the recognition process on the digital counting machine [In Russian]. Biophysics. – 1961. – V. 4, No. 2. – P. 17-23.
[7] ***Braverman EM.*** On the method of potential functions [In Russian]. Automation and Remote Control. – 1965. – V. 26, No. 12. – P. 2205-2213.
[8] ***Aizerman MA., Braverman EM., Rozonoer LI.*** Theoretical fundamentals of the method of potential functions in the problem of teaching machines to split the input into classes of situations [In Russian]. Automation and Remote Control. – 1964. – V. 25, No. 6. – P. 917-936.
[9] ***Borgest NM.*** Pattern recognition in designing artifacts is as metaphor and as an applied technology of ontology of designing [In Russian] *Ontology of designing.* 2015; 5(1): 19-29.
[10] ***Simankov VS., Lutsenko EV.*** Adaptive Control of Complex Systems Based on the Theory of Pattern Recognition [In Russian]. – Krasnodar: Kuban. Tekhnol. Univ, 1999. – 318 p.
[11] ***Kharkevich AA.*** On the choice of features for machine identification [In Russian]. Izv. Acad. of Sci. of the USSR, ser. tech. cybern. – 1963. – No. 2. – P. 3-9.
[12] ***Kanal LN., Chandrasekaran B.*** On dimensionality and sample size in statistical pattern classification. Pattern Recogn. – 1971. – V. 3, No. 3. – P. 225-234.
[13] ***Vapnik VN., Chervonenkis AYa.*** On a class of algorithms for pattern recognition learning [In Russian]. Automation and Remote Control. – 1964. – V. 25, No. 6. – P. 937-945.
[14] ***Babu KCh, Kalra SN.*** About the use of potential functions Bashkirov, Braverman and Muchnik for the selection of informative features in pattern recognition [In Russian]. Automation and Remote Control. – 1972. – V. 33, No. 12. – P. 105-107.

[15] ***Smirnov N.V.*** On the approximation of densities of distribution of random variables [In Russian]. Scientific Notes VP. Potemkin's MGPI. – 1951. – V. 16, No. III. – P. 69-96.
[16] ***Slutsky EE.*** Selected Works: Theory of Probability. Mathematical Statistics [In Russian]. – Moscow: Acad. of Sci. of the USSR Press, 1960. – 291 p.
[17] ***Vapnik VN., Chervonenkis AYa.*** On uniform convergence of frequencies of occurrence of events to their likely professions [In Russian]. Dokl. Acad. of Sci. of the USSR (DAS). – 1968. – V. 181, No. 4. – P. 781-784.
[18] ***Nilsson N.*** Learning Machines: Foundations of Trainable Pattern-Classifying Systems. – N. Y.: McGraw-Hill, 1965. 137 p.
[19] ***Fleishman BS.*** Elements of the Theory of Potential Effectiveness of Complex Systems [In Russian]. – Moscow: Sov. Radio, 1971. – 224 p.
[20] ***Lecky EK.*** Optimization solutions for the design of network data collection [In Russian]. – Vestn. MIIT. – 1998. – V. 1. – P. 125-130.
[21] ***Rozenberg GS., Mirkin BM., Ruderman YuS.*** Experience in applications of the theory of pattern recognition for the evaluation of soil salinity on vegetation [In Russian]. – Rus. J. Ecology. – 1972. – No. 6. – P. 31-34.
[22] ***Rozenberg GS.*** Reduction of the number of features and the evaluation of the soils for vegetation, when using methods of pattern recognition [In Russian]. – Quantitative Methods of Vegetation Analysis: Proceedings of the IV All-Union Conference "The Application of Quantitative Methods in the Analysis of Vegetation Structure". – Ufa: Acad. of Sci. of the USSR Press, 1974. – P. 36-39.
[23] ***Vasilevich VI.*** Second meeting "Application of quantitative methods in the study of the research Institute structure of vegetation" (Tartu, 1969) [In Russian]. – Bot. J. – 1970. – V. 55, No. 1. – P. 140-142.
[24] ***Mirkin BM.*** The Review "Theoretical Issues of Phytoindication" [In Russian]. – Bot. J. – 1972. – V. 57, No. 12. – P. 1342-1344.
[25] ***Shitikov VK., Rozenberg GS., Zinchenko TD.*** Quantitative Hydroecology: Methods, Criteria, Solutions: 2nd Vol. [In Russian]. – Moscow: Nauka, 2005. – V. 1. 281 p.; V. 2. 337 p.
[26] ***Vittikh VA.*** Paradigm of Bounded Rationality Decision-Making: Preprint [In Russian]. – Samara: ICSCS RAS Press, 2009. – 26 p.
[27] ***Smirnov SV.*** Ontological analysis of the subject areas modeling [In Russian]. – Izv. Samar. SC RAS. – 2001. – V. 3, No. 1. – P. 62-70.
[28] ***Smirnov SV.*** Ontological modeling in situational management [In Russian]. – Ontology of Designing. – 2012. – No. 2 (4). – P. 16-24.
[29] ***Rozenberg GS.*** Models in Phytocenology [In Russian]. – Moscow: Nauka, 1984. – 240 p.
[30] ***Rozenberg GS.*** Introduction to Theoretical Ecology. 2nd Vol., Ed. 2nd. [In Russian]. – Toglyatti: Kassandra, 2013. – V. 1. 565 p.; V. 2. 445 p.
[31] ***Rozenberg GS.*** On simple, complex, and supercomplex systems [In Russian]. – Fourth International Scientific-Practical Conference "Open Evaluation Systems" (May 20-21, 2016). Proceedings: Part 1. – Nizhyn (Ukraine): Univ. Press, 2016. – P. 228-233.

Сведения об авторе



Розенберг Геннадий Самуилович, 1949 г. рождения. Окончил Башкирский государственный университет (физико-математический и биологический факультеты) в 1966 г. Д.б.н., профессор, член-корреспондент РАН (с 2000 г.). Директор Института экологии Волжского бассейна РАН (с1989 г.). Опубликовал более 800 научных работ (в том числе более 60 монографий) в области теоретической экологии и моделирования экосистем.

Gennady S. Rozenberg (b. 1949). Graduated from the Bashkir State University (physic-mathematical and biological faculties; 1966). Doctor of biological Sciences, Professor, corresponding member of the Russian Academy of Sciences (2000). Director of the Institute of Ecology of the Volga River Basin of the RAS (1989). He has published more than 800 scientific papers (including more than 60 monographs) in the field of theoretical ecology and ecosystem modeling.