УДК 681.3

# ПРАКТИЧЕСКИЕ АСПЕКТЫ ПРИМЕНЕНИЯ КЛАССИФИКАТОРА СО СЛУЧАЙНЫМИ ПОДПРОСТРАНСТВАМИ

*Д.В. Жора*

Институт программных систем НАН Украины
03680, Киев, проспект Академика Глушкова, 40.
Тел.: 526 1538, факс 526 6263, email: dvz73@bigfoot.com

Актуальное современное направление, связанное с извлечением знаний из данных, во многом является применением методов классификации и распознавания образов. Традиционная постановка задачи классификации предполагает представление данных в виде множества вещественных векторов. В то же время, для многих практических задач такая постановка не является адекватной. В данной работе рассматривается применение классификатора со случайными подпространствами для решения задач с неполными данными и категориальными атрибутами. Предлагаются алгоритмы кластеризации и распределенного анализа данных.

The data mining algorithms is a modern topic in the area of computational intelligence. However, many solutions are based on well-known methods of classification and pattern recognition. For traditional classification task the data are represented as the set of real-valued vectors. At the same time, such approach is not suitable for many practical tasks. This article analyzes the application of random subspace classifier for datasets with missing values and categorical attributes. The clustering and distributed data processing algorithms are suggested.

## Области применения методов классификации и разпознавания образов

Многие ранние работы по распознаванию образов были ориентированы на воспроизведение функций выполняемых человеком, к которым в первую очередь следует отнести анализ изображений и распознавание речи. Среди популярных задач этого направления следует выделить распознавание символов (optical character recognition), распознавание текстур, распознавание лиц. На текущий момент эти задачи успешно решаются, при этом, обобщающая способность большинства систем существенно уступает возможностям человеческого восприятия. С другой стороны, проблема распознавания речи так и осталась нерешенной в предположении качественного представления речевой информации и независимости результатов от предъявленного диктора. Решение этой задачи позволило бы эффективно автоматизировать ввод неструктурированной текстовой информации в вычислительную среду и имело бы большое экономическое значение.

Современные применения методов классификации обусловлены появлением большого количества корпоративных баз данных. Представленные выборки данных могут быть использованы для обучения систем прогнозирования и анализа информации. По существу, системы классификации и извлечения знаний из данных [1] выполняют одни и те же функции, различие состоит только в нацеленности на разный конечный результат. Методы классификации ориентированы на повышение обобщающей способности системы, однако описание поведения системы, как правило, является достаточно сложным. Алгоритмы извлечения знаний из данных ориентированы на предоставление относительно простого для восприятия результата и аргументированное принятие решений, соответственно, качество прогнозирования или анализа данных не является оптимальным. Характерными представителями первого класса систем являются нейронные сети и метод опорных векторов. Эффективное описание областей принадлежности к классам как правило предоставляется деревьями решений.

Существует множество применений этих технологий в бизнесе, например, анализ или кластеризация посетителей интернет-портала, выделение наиболее перспективных клиентов интернет-магазина, обнаружение фальсификации данных при выдаче кредита, выявление товаров с коррелирующими временными паттернами продаж и т.д. Как правило, для большинства задач данной сферы выборки необходимые для обучения системы уже имеются в наличии. С другой стороны, при постановке задачи множество атрибутов, предоставляющих полезную информацию может быть расширено. Результаты таких исследований позволяют избежать ошибок в принятии решений, минимизировать финансовые риски, целенаправленно построить рекламную компанию, увеличить эффективность бизнеса благодаря учету потребностей клиента и т.п.

Другое направление связано с обеспечением безопасности передвижения людей и грузов. Рассмотрим, например, случай пассажирских авиаперевозок. Так или иначе, каждому пассажиру авиарейса можно поставить

в соответствие некоторый набор атрибутов, который его характеризует – возраст, пол, тип визы, гражданство, национальность, стоимость билета, средство оплаты билета, число дней от покупки билета до вылета, аэропорт отправления, аэропорт прибытия и т.д. Несмотря на вмешательство в частную жизнь граждан, многие страны стремятся расширить этот список. К категории опасных следует отнести пассажиров, которые осуществляли противозаконные действия, проносили на борт судна запрещенные предметы или вещества. Таким образом, данные по прошедшим перевозкам могут использоваться в качестве обучающей выборки системы. Применение такой системы позволяет производить более тщательный контроль лиц, относимых алгортмом классификации к категории потенциально опасных.

Большинство коммерческих грузов пересекают границу того или иного государства в контейнерах. При этом, тщательной проверке подвергается лишь малая доля контейнеров. Соответственно, вероятность проверки, как правило, является секретной информацией. Грузовой контейнер может содержать взрывчатку, наркотики, оружие, биологически опасные микроорганизмы и т.д. Каждому контейнеру можно поставить в соответствие следующий набор атрибутов: вес, характер груза, пункт отправления, пункт назначения, страна-отправитель, страна-производитель, число дней в пути, тип транспорта при пересечении границы и т.д. Система классификации позволяет проводить не случайную выборочную проверку грузов, а целенаправленый отбор и проверку потенциально опасных контейнеров.

Перспективным направлением применения методов классификации является задача прогнозирования эффективности лекарственных препаратов по соответствующему набору биологически активных компонент. Дело в том, что клинические испытания препаратов являются сравнительно дорогостоящими, занимают время от нескольких месяцев до нескольких лет и требуют участия достаточного количества пациентов. С другой стороны, число лекарств, которые могут быть построены из некоторого фиксированного набора компонент с заранее известными свойствами, является очень большим. Обучающей выборкой в данном случае является представление эффективности известных лекарственных средств в зависимости от набора составляющих их компонент. Таким образом, результатом применения методов искусственного интелекта является эффективный отбор наиболее перспективных соединений для проведения клинических испытаний.

Многие современные компании имеют достаточно большие корпоративные базы данных документов, число которых может исчисляться сотнями тысяч. Соответственно, задача поиска требуемой информации является актуальной и сложной, в контексте того, что текстовая информация является неструктурированной и не может быть непосредственно использована некоторым алгоритмом классификации. При этом, исходное множество документов может быть разделено на требуемое число тем или категорий автоматически при использовании методов предобработки текстовой информации в структурированное представление. Одна из возможностей реализации данного преобразования заключается в определении наиболее информативных слов, характеризующих некоторую категорию документов. Критерием полезности терма может являться количество информации (information gain) по отношению к индикатору правильной категории, которое эмпирически оценивается на заданном множестве документов. Таким образом, каждый текстовый документ может быть представлен бинарным вектором, характеризующим вхождение наиболее информативных термов для каждой из категорий. Соответственно, система извлечения знаний из данных может обучаться на существующих представлениях, а затем быть использована для автоматической классификации новых документов.

## Описание классификатора со случайными подпространствами

Данный раздел работы представляет алгоритм функционирования нейросетевого классификатора со случайными подпространствами. Более детальное описание математической модели сети приведено, например, в исследованиях [2, 3]. В частности, проведен вероятностный анализ преобразования входной информации, аналитически получены условия эффективной работы сети в зависимости от ее конфигурации и распределения входных данных. Продемонстрирована возможность распараллеливания программной реализации сети для использования на многопроцессорной либо многоядерной архитектуре [4]. Обобщающая способность сети была оценена на достаточно репрезентативной классификационной базе данных [5]. Были проанализированы основные преимущества используемой схемы грубого кодирования входных векторов [6].

Классификатор, общая схема которого приведена на рис. 1, состоит из четырех слоев нейронов: входного слоя пороговых элементов $\mathbf{l}_{ij}$ и $\mathbf{h}_{ij}$, слоев $\mathbf{A}$, $\mathbf{B}$ и $\mathbf{C}$. Первые три слоя осуществляют нелинейное

преобразование вещественного входного вектора в бинарный вектор большой размерности, который представляется слоем **B**. Схема такого преобразования показана на рис. 2. Последние два слоя **B** и **C** представляют собой обычный однослойный персептрон с матрицей связей **W**.


Рис. 1. Схема преобразования входного вещественного пространства

Архитектура нейронной сети зависит от следующих параметров: $N$ – количество нейронных групп $\mathbf{G}_j$ или нейронов слоя **B**; параметр подпространств $\eta$, который удовлетворяет ограничению $1 \le \eta \le n$, где $n$ – размерность входного пространства; параметры рецептивного поля $\delta_1$ и $\delta_2$, которые удовлетворяют условиям $0 < \delta_2$, $0 \le \delta_1 \le \delta_2$. Перед обучением сети должна быть сгенерирована ее структура, которая включает пороговые значения $l_{ij}$, $h_{ij}$ и индекс подпространств $\varphi(i, j)$, $i = \overline{1,\eta}$, $j = \overline{1,N}$.

Значения порогов вычисляются по формулам $l_{ij} = \xi_{ij} - \zeta_{ij}$ и $h_{ij} = \xi_{ij} + \zeta_{ij}$, где центр рецептивного поля $\xi_{ij}$ – случайная величина с равномерным распределение на отрезке $[-\delta_2, 1+\delta_2]$, полуширина поля $\zeta_{ij}$ – случайная величина с равномерным распределением на $[\delta_1, \delta_2]$. Каждая компонента $\varphi(i, j)$ является дискретной случайной величиной, которая принимает с равной вероятностью значения $\overline{1,n}$. На практике используются различные значения $\varphi$ в пределах одной нейронной группы, т.е. для некоторого индекса $j$.

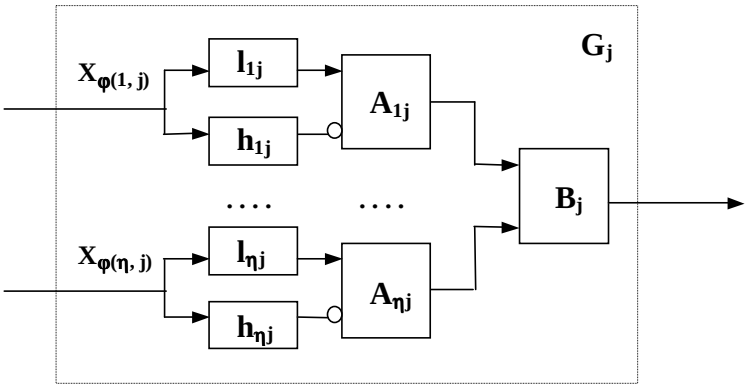


Рис. 2. Схема рецептивной группы классификатора со случайными подпространствами

Прямое распространение информации в сети осуществляется следующим образом. Нейрон $\mathbf{B}_j$, который может иметь выходной сигнал 0 или 1, возбуждается если $\forall i \in \overline{1,\eta}: l_{ij} < x_{\varphi(i,j)} < h_{ij}$. Фактически, нейрон $\mathbf{B}_j$ реагирует на попадание точки входного пространства в область, которая ограничена гиперплоскостями $x_{\varphi(i,j)} = l_{ij}$ и $x_{\varphi(i,j)} = h_{ij}$, а группа $\mathbf{G}_j$ функционирует в соответствующем линейном подпространстве. Постсинаптический потенциал нейронов слоя **C** вычисляется как $u_k = \sum_{j=1}^{N} W_{kj} B_j$. Нейрон с наибольшим потенциалом и определяет класс, к которому классификатор относит входной вектор.

Обучение сети осуществляется по правилу с фиксированными приращениями, которое было предложено Розенблаттом для однослойного персептрона [7]. Изменение синапсов $W_{ij}$ производится в случае ошибочной классификации. Пусть $t$ это исходный номер класса, а $f$ – номер класса, который получен классификатором. Тогда синаптические коэффициенты пересчитываются по формулам $W'_{tj} = W_{tj} + B_j$ и $W'_{fj} = \max(W_{fj} - B_j, 0)$. Иногда используется симметричное правило, когда $W'_{fj} = W_{fj} - B_j$.

Любая конечная задача классификации может быть приведена в единичный гиперкуб с помощью линейного преобразования. Теоретический анализ работы сети, как правило, предполагает, что все входные векторы являются нормализованными. Следует отметить некоторые практически значимые результаты аналитического исследования данной модели [3]. В частности, линейная плотность активных пороговых гиперплоскостей, пересечение которых приводит к изменению бинарного образа входного вектора, определяется по следующей формуле:

$$g = \frac{2\eta\lambda^{\eta} N}{n(\delta_2 + \delta_1)}, \text{ где } \lambda = \frac{\delta_2 + \delta_1}{1 + 2\delta_2}.$$

Вероятность неразличимости двух точек единичного гиперкуба $\mathbf{x}$ и $\mathbf{y}$ оценивается выражением $\exp(-g \cdot z)$, где $z = \sum_{i=1}^{n} | x_i - y_i |$ – блочное расстояние между точками. Классификатор со случайными подпространствами является универсальным классификатором. Другими словами, при использовании стандартной процедуры обучения сети практически любая обучающая выборка может быть проинтерпретирована без ошибок. При этом, данное свойство не имеет непосредственной импликации на обобщающую способность сети.

## Алгоритм решения задач классификации с неполными данными

Существенным аспектом практической значимости методов классификации является возможность их применения для решения задач с неполными входными данными. Другими словами, некоторые компоненты входных векторов могут быть неопределенными, т.е. соответствовать значению NULL в терминах баз данных. Традиционный метод решения этой проблемы заключается в подстановке некоторых реальных значений на место неопределенных компонент, возможно, с учетом эмпирического вероятностного распределения. При этом, такой подход обладает тем недостатком, что способствует искажению исходного вероятностного распределения входных данных В данном разделе приведен простой, но достаточно эффективный алгоритм решения задач классификации с неполными данными для классификатора со случайными подпространствами, который заключается в стохастическом поведении системы.

В случае, когда все входные значения определены, после генерации структуры сети классификатор со случайными подпространствами следует рассматривать как детерминированный автомат. Если же на вход элементов $\mathbf{l}_{ij}$ и $\mathbf{h}_{ij}$ с фиксированными пороговыми значениями подается неопределенная компонента, то выходное значение нейрона $\mathbf{A}_{ij}$ должно быть независимой бинарной случайной величиной с вероятностью активации равной длине пересечения отрезков $[l_{ij}, h_{ij}]$ и $[0,1]$. Тогда значение $\max(0, \min(h_{ij}, 1) - \max(0, l_{ij}))$ соответствует вероятности активации нейрона $\mathbf{A}_{ij}$ и является оптимальным в предположении равномерного распределения компоненты на отрезке $[0,1]$. Приведенное правило следует применять как при обучении сети, так и в режиме экзамена. Как отмечалось, в данном случае классификатор со случайными подпространствами становится стохастическим автоматом.

Возможно обобщение данного метода на случай генерации чуствительной структуры классификатора, когда плотность пороговых значений становится пропорциональной плотности вероятности распределения соответствующих компонент входных данных [2, 5]. В этом случае, значения активации нейронов $\mathbf{A}_{ij}$ следует запоминать при генерации пороговых значений до преобразования входного пространства. Тогда, вероятность активации данного нейрона будет приблизительно равна вероятности попадания компоненты входного вектора между порогами $\mathbf{l}_{ij}$ и $\mathbf{h}_{ij}$ в соответствии с эмпирическим вероятностным распределением.

Применимость данного алгоритма была проверена на множестве задач классификации полученных с использованием алгоритма DataGen [8]. Рис. 3 демонстрирует зависимость уровня ошибок класификации от вероятности обращения некоторой компоненты входного вектора в значение NULL. Использовались следующие параметры эксперимента: число категорий $m = 2$, размерность входного пространства $n = 10$, размерность подпространств $\eta = 3$, параметры рецептивного поля $\delta_1 = \delta_2 = 0.5$, число рецептивных групп сети $N = 32768$, сложность алгоритма DataGen – 1, 5 и 9 для каждой представленной зависимости соответственно. Линейность полученных зависимостей объясняется, в основном, равномерностью распределения точек внутри единичного гиперкуба для решаемой задачи классификации.

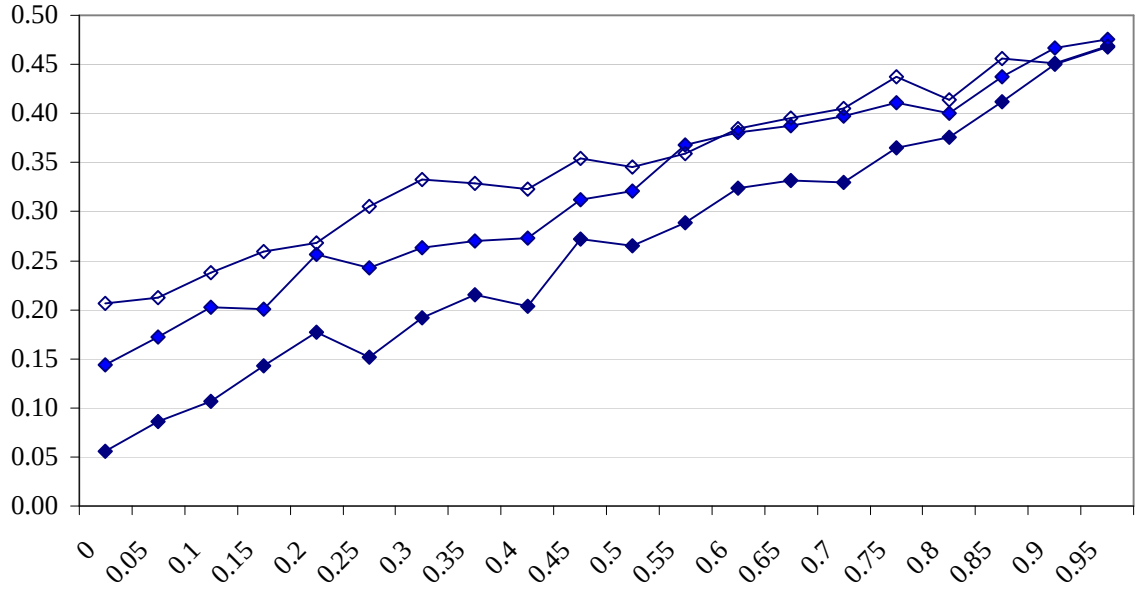


Рис. 3. Зависимость эмпирической вероятности ошибок от доли пропущенных компонент векторов

Следует отметить возможность кеширования бинарных образов обучающей выборки в представлении слоя **B** . Обобщающая способность сети оказывается лучше в том случае, когда бинарные образы строятся и фиксируются напротяжении всего процесса обучения. Альтернативой является перегенерирование бинарных образов для каждой эпохи обучения. В последнем случае, стохастическое поведение алгоритма предобработки информации в классификаторе фактически привносит шумовые составляющие в процесс обучения сети.

Предпосылки и мотивация для решения задач с неполными данными могут быть самыми разными. В частности, большую роль в предобработке данных играет человеческий фактор. Во многих случаях, наличие тех или иных данных определяется бизнес процессами, некоторые параметры могут иметь область определения не соответствующую множеству всех возможных исходов. Как показано выше, такие задачи классификации могут успешно решаться с получением практически полезных результатов.

## Алгоритм учета практической стоимости ошибок классификации

Рассмотрим задачу анализа данных при посадке пассажиров авиалайнера. Очевидно, что стоимость ошибки, которая заключается в пропуске преступника в салон, несравнимо больше, чем тщательный досмотр обычного пассажира и вероятные потери авиакомпании за счет выбора пассажирами других видов транспорта. Соответственно, качество обнаружения угроз должно быть значительно выше, а реакция системы должна быть существенно “подозрительнее” именно за счет большой стоимости ошибки.

Аналогичная ситуация имеет место при классификации заявок на выдачу кредита. В данном случае количество документов, которые содержат фальсифицированные данные, значительно меньше общего числа заявок. Например, обучающая выборка данных может содержать только 10 % векторов первой категории и 90 % векторов, соответствующих корректно подготовленным документам. При таком представительстве данных большинство систем прогнозирования, в том числе и классификатор со случайными подпространствами, будут иметь тенденцию поглощения меньшего класса. В результате, при тестировании порядка 99 % векторов может быть отнесено ко второй категории.

Предположим, что каждая из категорий должна распознаваться с одинаковым качеством. Это означает, что следует минимизировать не средний уровень ошибок, а сбалансированный уровень ошибок (balanced error rate), который в данном случае вычисляется по следующей формуле:

$$BER = \frac{F_{12}}{2(F_{11} + F_{12})} + \frac{F_{21}}{2(F_{21} + F_{22})},$$

где $F_{tf}$ – число классификаций на тестовой выборке класса $f$ , когда на самом деле вектор относится к классу $t$ . Значит, для увеличения вероятности обнаружения фальсификации первая категория должна иметь большее подкрепление, что в свою очередь эквивалентно большей стоимости ошибки классификации.

Рассмотрим более общий случай классификации на $m$ категорий $C_k$ , где $k = \overline{1,m}$ . Для большинства прикладных задач следующая модель учета неравноценности ошибок является достаточно адекватной. Задается

матрица стоимостей ошибок классификации (classification cost matrix) $\mathbf{D}$ размерностью $m \times m$, где значение $D_{tf}$ – стоимость ошибки определения класса $f$, когда на самом деле вектор относится к классу $t$. Матрица ошибок называется простой когда все диагональные элементы равны нулю, кроме того, все недиагональные элементы равны между собой и больше нуля. Для простой матрицы оптимизация стоимости классификации эквивалентна минимизации среднего уровня ошибок. В общем случае компоненты матрицы стоимостей ошибок могут быть отрицательными, что соответствует получению пользы или прибыли от совершения ошибки.

Алгоритм обучения классификатора со случайными подпространствами с учетом стоимости ошибок распознавания предполагает вещественнозначные синаптические веса и заключается в следующем. В случае ошибочной классификации некоторого вектора из обучающего множества веса синаптической матрицы пересчитываются по формулам $W'_{tj} = W_{tj} + B_j D_{tf}$ и $W'_{fj} = \max(W_{fj} - B_j D_{ft}, 0)$. Может также использоваться правило без условия неотрицательности: $W'_{fj} = W_{fj} - B_j D_{ft}$.

Таким образом, нейрон слоя $\mathbf{C}$, соответствующий категории с большой стоимостью ошибки, получает большее подкрепление. Соответственно, такая категория будет иметь большую вероятность быть выбранной классификатором. Приведенный алгоритм соответствует концепции обратного распространения ошибки (back propagation) для персептрона [7] и имеет некоторые аналогии с методами бустинга (boosting) [9]. Качество данного алгоритма требует экспериментальной оценки с точки зрения минимизации суммарной стоимости классификации на тестовой выборке данных, так как существенно зависит от вероятностных распределений решаемой задачи.

Возможен вариант разрешения проблемы неравномерного представительства классов в обучающей выборке с использованием классической модели классификатора. Для этого достаточно расширить обучающую выборку за счет дублирования векторов для категорий с малой представительностью. При этом, полученное число векторов для каждой категории должно быть приблизительно одинаковым. Если $L$ – обшая длина обучающей выборки с разбиением по категориям $L_k$, где $k = \overline{1,m}$, то суммарная длина расширенной выборки должна составлять $m \cdot \max\limits_k L_k$.

## Адаптация архитектуры сети для анализа категориальных атрибутов

До настоящего времени в качестве компонент входных векторов рассматривались вещественные значения. С другой стороны, в современных корпоративных базах данных достаточно часто встречаются атрибуты, которые имеют конечное дискретное множество значений и не имеют каких-либо отношений порядка. Будем называть такие атрибуты категориальными (categorical attributes), подразумевая, что значение атрибута может быть поставлено в однозначное соответствие некоторой категории.

В качестве примера таких атрибутов можно привести цвет – множество значений может состоять из элементов “красный”, “синий” и “зеленый”. В задаче анализа запросов к интернет серверу достаточно полезная информация может быть представлена в виде атрибута страны, из которой исходит запрос. Частным случаем категориальных атрибутов являются бинарные атрибуты. С другой стороны, бинарные атрибуты являются частным случаем вещественных. В представлении баз данных категориальные атрибуты, как правило, имеют строковый тип, хотя целочисленное представление тоже возможно.

В общем случае, схема преобразования входного пространства для классификатора со случайными подпространствами должна обеспечивать на выходе бинарное представление для множества вещественных и категориальных атрибутов. Схему выбора подпространств следует оставить без изменения – каждому нейрону $\mathbf{A}_{ij}$ ставится в соответствие случайно выбранное измерение, которое может быть либо вещественным, либо категориальным. Рассмотрим далее случай категориального измерения.

Вместо пары пороговых значений бинарному нейрону $\mathbf{A}_{ij}$ должно быть поставлено в соответствие случайно выбранное значение категориального атрибута $a_{ij}$. При равномерной структуре классификатора требуемое значение выбирается равновероятно из множества всех значений по данному измерению. Если производится генерация чувствительной структуры сети [2], то вероятность выбора значения $a_{ij}$ должна быть равна эмпирической вероятности появления данного значения для соответствующего измерения.

Нейрон $\mathbf{A}_{ij}$ возбуждается в том случае, если на его вход подается значение равное $a_{ij}$. В остальных случаях выходной сигнал нейрона должен быть равен нулю. Дальнейшее распространение информации в сети осуществляется по стандартным правилам, представленным во втором разделе работы. В частности, нейрон $\mathbf{B}_j$ возбуждается в случае когда все афферентные к нему нейроны $\mathbf{A}_{ij}$ активизированы.

Следует выделить следующие аспекты применения данного алгоритма. В случае большого количества значений некоторого категориального атрибута вероятность возбуждения соответствующих нейронов слоя $\mathbf{A}$ будет сравнительно малой. Следовательно, плотность бинарного кодирования в слое $\mathbf{B}$ будет меньше. Кроме

того, плотность кодирования может быть неравномерной в зависимости от индекса $j$, так как нейрон $\mathbf{B}_j$ будет иметь меньшую вероятность активации, если некоторый из его афферентных нейронов функционирует по категориальному измерению с большим количеством значений. С другой стороны, подобная проблема будет присуща практически любому методу классификации. По аналогии с проблемой большой размерности простых средств устранения указанных недостатков не существует.

Рассмотрим случай предъявления неопределенного значения категориального атрибута. Пусть для определенности данный атрибут принимает значения $a_k$ с вероятностями $p_k$, $k=\overline{1,K}$. Если классификатор имеет равномерную структуру, то выходное значение элемента $\mathbf{A}_{ij}$ должно быть независимой бинарной случайной величиной с вероятностью активации равной $1/K$. В случае чуствительной структуры вероятность активации должна быть равной $a_k$, где индекс $k$ определяется из соотношения $a_k = a_{ij}$. При программной реализации вероятности активации удобно кешировать во время генерации структуры классификатора.

## Метод выявления пространственного кластера входных данных

Основным отличием данной задачи от постановки задачи классификации является то, что обучающая выборка данных не содержит меток классов. Другими словами, все рассматриваемые точки относятся к одной и той же категории – к кластеру данных. Геометрически проблема заключается в отделении области, в которой расположена основная часть точек обучающей выборки, от пустого пространства.

Несмотря на отсутствие очевидного критерия ограничения размеров области кластера, алгоритмы решения этой задачи имеют применение в маркетинге, в рекламном бизнесе и т.п. Предположим, что вектора обучающей выборки содержат данные о текущих клиентах компании. Имея геометрическое описание формы кластера, либо алгоритм, который может произвести верификацию принадлежности точки к кластеру, можно дать ответы, например, на следующие вопросы. Является ли новый клиент компании типичным представителем их клиентской базы? Будет ли новый клиент заинтересован в долгосрочном сотрудничестве? Предпочтет ли он в дальнейшем воспользоваться услугами конкурентов, которые занимают другие ниши рынка? Как грамотно построить рекламную компанию, чтобы целевой аудиторией являлись потенциальные клиенты с такими же характеристиками как у существующих? Существует ли область кластера, которая не является целевой для текущей рекламной программы, но является перспективной для расширения бизнеса? Приведенные вопросы являются особенно актуальными для интернет-магазинов, где рыночный спрос и предложение изменяются во времени достаточно быстро.

Адаптация классификатора со случайными подпространствами для решения этой задачи может быть произведена следующим образом. Выходной слой $\mathbf{C}$ должен иметь два нейрона, первый нейрон соответствует категории кластера, а второй – пустому пространству. Обучение сети производится за некоторое количество эпох. Одной эпохой называется единичный проход обучающей выборки с обучением сети на каждом векторе данных. При предъявлении входного вектора первая фаза обучения сети полностью совпадает с алгоритмом, представленным во втором разделе. Таким образом, осуществляется подкрепление связей между бинарным представлением вектора и категорией кластера. С другой стороны, связи бинарного представления с нейроном пустого пространства ослабляются.

Значит, для некоторого входного вектора $\mathbf{x}$ его бинарное представление $\mathbf{p}$ вычисляется на первом этапе. Вторая фаза обучения заключается в подкреплении категории пустого пространства для бинарного дополнения представления входного вектора в слое $\mathbf{B}$. Другими словами, вычисляется побитовое отрицание вектора $\mathbf{p}$ и производится его случайное прореживание, чтобы плотность единиц в результирующем векторе $\mathbf{p}^*$ была сравнимой с плотностью исходного вектора $\mathbf{p}$. Если число единиц вектора $\mathbf{p}$ составляет $N_p$, то каждая компонента бинарного отрицания должна обнуляться со следующей вероятностью:

$$\alpha_p = \frac{c \cdot (N - 2N_p)}{(N - N_p)},$$

здесь $c$ – параметр открытости кластера, по умолчанию равный единице. При увеличении коэффициента $c$ линейные размеры кластера тоже увеличиваются. Для полученного вектора $\mathbf{p}^*$ производится стандартная процедура модификации синаптических коэффициентов по формулам $W'_{2j} = W_{2j} + p^*_j$ и $W'_{1j} = \max(W_{1j} - p^*_j, 0)$. Возможно использование правила $W'_{1j} = W_{1j} - p^*_j$.

В режиме полного обучения алгоритм останавливается, если распределение точек по категориям не изменяется на протяжении нескольких эпох, число которых предопределяется заранее. Существенной также является возможность перекрестной проверки (cross-validation). Предположим наличие дополнительного независимого множества кластерных точек с мощностью $L_I$. Проверочная выборка получается добавлением к данному множеству $[L_I/c]$ случайных векторов с равномерным распределением внутри единичного гиперкуба

и метками класса, соответствующими пустому пространству. Ранняя остановка обучения должна происходить при достижении минимума ошибки на полученной проверочной выборке.

## Решение задачи кластеризации с использованием классификатора

Как задача определения кластера входных данных так и задача кластеризации на несколько категорий относятся к классу автономного обучения (unsupervised learning). Фактически, это означает, что в систему не вносится никакая дополнительная информация, кроме обучающей выборки данных. Контролируемое обучение (supervised learning) в случае классификации отличается тем, что на каждом шаге процесса обучения в систему передается информация о метке класса для соответствующего вектора. При этом, количество категорий для проблемы кластеризации тоже является предопределенным заранее.

Задачи кластеризации, как правило, возникают при анализе больших объемов данных в некоторой новой области исследования. Определение первоначального количества кластеров обычно носит эвристический характер. Полученное решение дает предварительную информацию о структуре данных, о свойствах кластеров и основных характеристиках объектов. В дальнейшем проблема может быть переформулирована в задачу классификации после спецификации требуемых категорий и характеристик. Важным применением методов кластеризации является проблема кластеризации документов, так как для достаточно большой базы данных спецификация категорий не является очевидной или известной априори.

Для решения задачи кластеризации на $m$ категорий изменение архитектуры для классификатора со случайными подпространствами не является необходимым. С другой стороны, перед проведением процедуры обучения нужно внести в синаптическую матрицу сети бинарный шум. Другими словами, каждая компонента $W_{kj}$ должна независимо инициализироваться значением 0 или 1 с вероятностью $1/2$. Использование более сложного распределения значений не является мотивированным, так как каждый входной вектор может либо принадлежать либо не принадлежать некоторой категории. Данная нейронная сеть переводит близкие точки во входном пространстве в близкие бинарные представления в терминах расстояния Хемминга [2,3]. В случае вещественного входного пространства естественной метрикой сети во входном пространстве является блочная или манхеттенская метрика. Таким образом, сразу после внесения синаптического шума близкие бинарные представления в слое **B** будут имеют большую вероятность быть отнесенными к одной категории, а данная нейронная сеть будет обладать некоторыми кластеризационными свойствами.

Дальнейшее обучение кластеризатора следует производить в соответствии со следующим алгоритмом. Для каждого вектора обучающей выборки нейронная сеть определяет его категорию и производит операцию изменения синаптической матрицы в соответствии с формулой $W'_{tj} = W_{tj} + B_j$, где индекс $t$ соответствует полученной метке кластера. Процедура обучения обычно повторяется некоторое число эпох до тех пор, пока распределение векторов выборки по категориям изменяется. Методы ранней остановки обучения в данном случае рассматриваться не могут, так как кластеризация всегда является субъективной с точки зрения выбора метрики близости точек входного пространства. Вероятностных критериев оценки качества полученного решения также не существует.

Использование решения задачи кластеризации может быть различным. Полученные кластеры данных могут быть описаны геометрически – возможно вычисление таких показателей как центр кластера, размеры кластера по различным измерениям, дисперсионные характеристики. Области принадлежности к кластерам могут быть визуализированы для анализа специалистами в соответствующей прикладной области. Кластеры с малым числом векторов могут быть отброшены из рассмотрения либо объединены в один дополнительный кластер. Как было отмечено ранее, результаты кластеризации часто используются для формулировки задачи классификации.

## Анализ информации из распределенных баз данных

Большинство современных средств извлечения знаний из данных предполагает, что все необходимые выборки собраны в одной централизированной базе данных. В то же время, такое ограничение значительно сужает потенциальный рынок применения технологий искусственного интеллекта. Рассмотрим случай, когда две разные компании владеют различной информацией, например, по пересекающемуся множеству клиентов. Компании заинтересованы провести маркетинговое исследование и использовать для этого общие данные, но не могут передавать друг другу персональную информацию по своим клиентам.

Аналогично, передача частной информации между различными государственными ведомствами, как правило, ограничена законодательством. С другой стороны, анализ таких данных позволил бы сформулировать превентивные меры по снижению преступности, лучше противостоять угрозам общественной безопасности, упростить сотрудничество государственных и коммерческих организаций. Таким образом, есть потребность в алгоритмах анализа данных, которые бы не нарушали ограничений, связанных с владением конфиденциальной информацией, и не ущемляли бы коституционные права граждан.

Пусть распределенная система состоит из $M$ баз данных, каждая из которых содержит таблицу с полем уникального идентификатора и некоторыми атрибутами требуемого типа объектов или сущностей. Обозначим количество атрибутов, локально доступных для обработки на каждом компьютере как $n_\omega$, а соответствующие множества как $\nu_\omega$, здесь $\omega = \overline{1, M}$. Далее, пусть множество всех атрибутов обозначается как $\nu$. Очевидно, что распределение атрибутов по базам данных удовлетворяет следующему соотношению:

$$\sum_{\omega=1}^{M} n_\omega \geq n .$$

Предположим, что количество идентификаторов, которые являются общими для всех $M$ баз данных, составляет $L$. При этом, количество локальных записей таблицы для каждого компьютера может быть большим. Очевидно, что преобразование входных векторов в бинарное представление для классификатора со случайными подпространствами осуществляется независимо для каждой рецептивной группы. Разделим общее множество $N$ рецептивных групп на подмножества с мощностями $N_\omega$, где $\omega = \overline{1, M}$. Кроме того, имеет смысл потребовать, чтобы соотношение значений $n_\omega / N_\omega$ было приблизительно постоянным. Пусть подмножество с индексом $\omega$ может иметь входные параметры только из $\nu_\omega$. Тогда преобразование входного пространства в бинарное представление может осуществляться распределенно, а полученные бинарные подвекторы могут объединяться в общее представление с использованием сетевых коммуникационных технологий.

Программная архитектура распределенной системы состоит из одного центрального обработчика и $M$ локальных агентов. Каждый агент устанавливается локально по отношению к соответствующей базе данных. Функциональность агента заключается в генерировании локальной структуры подсети и получении бинарного представления для подмножества соответствующих рецептивных групп в режиме обучения или экзамена. Расположение центрального обработчика является произвольным. Реализация обработчика должна включать, как минимум, следующие функции: управление агентами в режиме генерации структуры, обучения и экзамена; выделение общего подмножества идентификаторов для всех локальных агентов системы; десериализация и агрегирование бинарных подпредставлений, получаемых от агентов; хранение и модификация синаптической матрицы; вычисление постсинаптического потенциала выходных нейронов и принятие решений.

Таким образом, предложенная архитектура системы предполагает передачу по сети только значений уникальных идентификаторов и бинарных подвекторов в представлении слоя **B**. Следовательно, возможность извлечения значений реальных атрибутов объектов является минимальной. Обобщающая способность сети может оказаться несколько ниже за счет исключения возможности комбинирования атрибутов, которые эксклюзивно расположены в разных базах данных. Тем не менее, принципиальная возможность извлечения знаний из распределенных баз данных существует, несмотря на значительные ограничения с точки зрения информационной безопасности.

## Заключение и анализ

Задачи классификации и кластеризации возникают в современном мире практически в каждой области деятельности человека – в естествознании, экономике, медицине и т.п. Можно утверждать, что классификатор со случайными подпространствами является достаточно удобным средством их решения. Данное исследование демонстрирует, что применимость классификатора со случайными подпространствами может быть значительно расширена за счет анализа реальной постановки задач распознавания, прогнозирования или классификации. Многие представленные методы могут быть использованы для других инструментов статистического обучения или методов извлечения знаний из данных.

В работе приведены алгоритмы работы нейронной сети с неполными данными и категориальными атрибутами. Показано, что модифицированный алгоритм обучения сети является более адекватным с точки зрения минимизации стоимости ошибок классификации. Рассматриваемый нейросетевой классификатор может быть использован для решения различных проблем кластеризации данных. Существенным фактором является то, что большинство рассмотренных подходов могут комбинироваться с учетом требований реальной задачи.

1. *Witten I.H., Frank E.* Data mining: practical machine learning tools and techniques. 2-nd edition, Elsevier, 2005, – 525 p.
2. *Жора Д.В.* Анализ функционирования классификатора со случайными порогами. Кибернетика и системный анализ. – 2003. – № 3. – С. 72-91.
3. *Жора Д.В.* Анализ формирования разделяющих поверхностей для классификатора со случайными подпространствами. Кибернетика и системный анализ. – 2006. – № 6. – С. 55-70.
4. *Жора Д.В.* Распараллеливание алгоритмов функционирования классификатора со случайными подпространствами // Проблемы программирования. – 2006. – № 2-3. – С. 124-134.
5. *Zhora D.V.*. Evaluating Performance of Random Subspace Classifier on ELENA Classification Database // Proc. Int. Conf. Artificial Neural Networks 2005, LNCS 3697, P. 343–349.
6. *Kussul E.M., Rachkovskij D.A., Wunsch D.C.* The random subspace coarse coding scheme for real-valued vectors. International Joint Conference on Neural Networks 1999. – Vol. 1. – P. 450–455.
7. *Haykin S.* Neural networks a comprehensive foundation. 2-nd edition, Upper Saddle River, NJ, Prentice Hall. – 1999. – 842 p.
8. *Rachkovskij D.A., Kussul E.M.* DataGen: a generator of datasets for evaluation of classification algorithms. Pattern Recognition Letters 19. – 1998. – P. 537–544.
9. *Duda R.O., Hart P.E., Stork D.G.* Pattern Classification. 2-nd edition, Wiley Interscience. 2000. – 654 p.
10. *Kussul E.M., Baidyk T.N., Lukovich V.V., Rachkovskij D.A.* Adaptive high performance classifier based on random threshold neurons. Cybernetics and Systems'94, Ed. R.Trappl, Singapore, World Scientific Publishing Co. Pte. Ltd. – 1994. – P. 1687–1695.