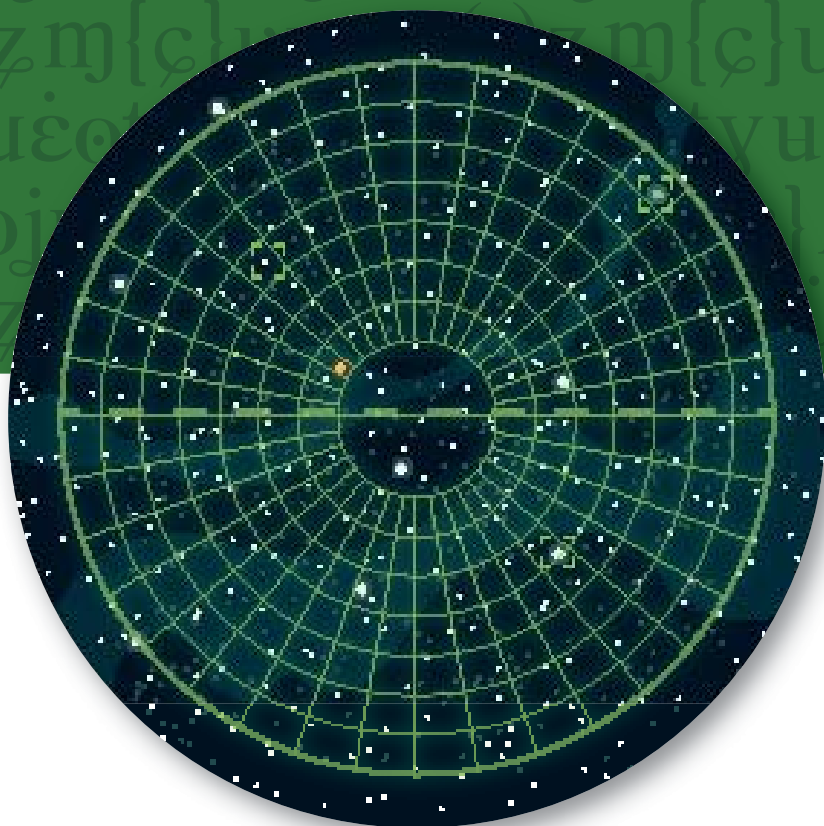




СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
SIBERIAN FEDERAL UNIVERSITY



Е. А. Сопов, И. А. Иванов
МНОГОКРИТЕРИАЛЬНЫЕ
НЕЙРОЭВОЛЮЦИОННЫЕ СИСТЕМЫ
В ЗАДАЧАХ МАШИННОГО ОБУЧЕНИЯ
И ЧЕЛОВЕКО-МАШИННОГО ВЗАИМОДЕЙСТВИЯ

УДК 004.032.26:378.147+004.5

ББК 32.818.1

C645

Р е ц е н з е н т ы:

Е. С. Семенкин, доктор технических наук, профессор кафедры системного анализа и исследования операций СибГУ имени академика М. Ф. Решетнева;

В. А. Терсков, доктор технических наук, профессор кафедры управления персоналом Красноярского института Железнодорожного транспорта – филиала ФГБОУ ВО ИрГУПС

Сопов, Е. А.

C645 Многокритериальные нейроэволюционные системы в задачах машинного обучения и человеко-машинного взаимодействия : монография / Е. А. Сопов, И. А. Иванов. – Красноярск : Сиб. федер. ун-т, 2019. – 160 с.

ISBN 978-5-7638-3969-2

Рассмотрены методы и модели машинного обучения для построения автоматизированных систем человеко-машинного взаимодействия.

Предназначена для студентов, аспирантов и научных работников, интересующихся проблемами проектирования методов и моделей машинного и глубинного обучения.

Электронный вариант издания см.:
<http://catalog.sfu-kras.ru>

УДК 004.032.26:378.147+004.5
ББК 32.818.1

ISBN 978-5-7638-3969-2

© Сибирский федеральный университет, 2019

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ	6
Глава 1. СИСТЕМНЫЙ АНАЛИЗ ПРОБЛЕМЫ ПРИМЕНЕНИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ И ОПТИМИЗАЦИИ В ЗАДАЧАХ ЧЕЛОВЕКО-МАШИННОГО ВЗАИМОДЕЙСТВИЯ	9
1.1. Обзор современных методов машинного обучения, классификации и оптимизации	9
1.2. Задача человеко-машинного взаимодействия и обзор существующих подходов к ее решению	29
1.3. Проблема распознавания эмоций при разработке человеко-машинных интерфейсов.....	36
Выводы по главе 1	40
Глава 2. КОЛЛЕКТИВНЫЙ САМОКОНФИГУРИРУЕМЫЙ ЭВОЛЮЦИОННЫЙ АЛГОРИТМ МНОГОКРИТЕРИАЛЬНОЙ ОПТИМИЗАЦИИ	42
2.1. Эволюционные алгоритмы однокритериальной и многокритериальной оптимизации	42
2.2. Разработка и реализация самоконфигурируемого эволюционного алгоритма многокритериальной оптимизации.....	54
2.3. Исследование эффективности самоконфигурируемого алгоритма на репрезентативном наборе тестовых задач оптимизации.....	58
Выводы по главе 2	66
Глава 3. МНОГОКРИТЕРИАЛЬНЫЙ ПОДХОД К ПРОЕКТИРОВАНИЮ АНСАМБЛЯ КЛАССИФИКАТОРОВ И ОТБОРУ ИНФОРМАТИВНЫХ ПРИЗНАКОВ	68
3.1. Настройка параметров и проектирование ансамблей алгоритмов машинного обучения.....	68
3.2. Разработка и реализация многокритериального подхода к отбору информативных признаков.....	74
3.3. Разработка и реализация многокритериального подхода к проектированию ансамбля нейросетевых классификаторов.....	81

3.4. Исследование эффективности многокритериального подхода к отбору информативных признаков и проектированию ансамбля нейросетевых классификаторов.....	86
Выводы по главе 3	97
 Глава 4. ГИБРИДНЫЙ АЛГОРИТМ ОБУЧЕНИЯ КОНВОЛЮЦИОННОЙ НЕЙРОННОЙ СЕТИ С ПРИМЕНЕНИЕМ ЭВОЛЮЦИОННОГО АЛГОРИТМА ОПТИМИЗАЦИИ.....	
4.1. Конволюционная нейронная сеть и суть методов глубинного обучения	99
4.2. Достоинства и недостатки алгоритмов обратного распространения ошибки и эволюционного алгоритма для настройки искусственных нейронных сетей	104
алгоритм обратного распространения ошибки	104
4.3. Гибридный алгоритм обучения конволюционной нейронной сети	108
4.4. Исследование эффективности гибридного алгоритма обучения конволюционной нейронной сети на задачах анализа изображений	112
Выводы по главе 4.....	117
 Глава 5. ОБОБЩЕННЫЙ МЕТОД ДЛЯ РЕШЕНИЯ ЗАДАЧ АНАЛИЗА ГЕТЕРОГЕННЫХ ДАННЫХ.....	
5.1. Метод слияния аудио-, видеоинформации на уровне данных и на уровне классификаторов в рамках задачи распознавания эмоций	119
5.2. Разработка обобщенного метода для решения задач анализа гетерогенных данных на основе слияния данных, многокритериального отбора признаков и оптимизации алгоритмов машинного обучения и конволюционных нейронных сетей.....	124
5.3. Исследование эффективности обобщенного метода на задаче распознавания эмоций	127
Выводы по главе 5	128
 ЗАКЛЮЧЕНИЕ	129
 СПИСОК ЛИТЕРАТУРЫ	131

Приложение А. СРАВНЕНИЕ РАЗРАБОТАННОГО АЛГОРИТМА <i>SELFCOMOGA</i> С АЛГОРИТМАМИ-ПОБЕДИТЕЛЯМИ СОРЕВНОВАНИЯ СЕС ПО МЕТРИКЕ <i>IGD</i>	143
--	-----

Приложение Б. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ ПО ИССЛЕДОВАНИЮ ЭФФЕКТИВНОСТИ МНОГОКРИТЕРИАЛЬНОГО ПОДХОДА К ОТБОРУ ИНФОРМАТИВНЫХ ПРИЗНАКОВ И ПРОЕКТИРОВАНИЮ АНСАМБЛЯ НЕЙРОСЕТЕВЫХ КЛАССИФИКАТОРОВ	146
--	-----

Приложение В. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ ПО ИССЛЕДОВАНИЮ ЭФФЕКТИВНОСТИ ГИБРИДНОГО АЛГОРИТМА ОБУЧЕНИЯ КОНВОЛЮЦИОННОЙ НЕЙРОННОЙ СЕТИ НА ЗАДАЧЕ РАСПОЗНАВАНИЯ ЭМОЦИЙ	153
--	-----

Приложение Г. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ ПО ИССЛЕДОВАНИЮ ЭФФЕКТИВНОСТИ ОБОБЩЕННОГО МЕТОДА АНАЛИЗА ГЕТЕРОГЕННЫХ ДАННЫХ НА ЗАДАЧЕ РАСПОЗНАВАНИЯ ЭМОЦИЙ	155
---	-----

ВВЕДЕНИЕ

Одна из задач машинного обучения заключается в построении модели по имеющейся базе данных в соответствии с некоторым алгоритмом. В общем случае алгоритмы машинного обучения не позволяют добиться высокой точности решения задачи без предварительной настройки параметров. Настройка параметров алгоритмов вручную может оказаться очень затратной по времени. Кроме того, эксперт в области машинного обучения должен обладать необходимыми знаниями о настраиваемом алгоритме и свойствах процесса обучения данного алгоритма.

Данная монография посвящена проблеме проектирования нейросетевых систем машинного обучения эволюционными алгоритмами при решении задач человеко-машинного взаимодействия.

Задача автоматизации проектирования методов и моделей машинного обучения не является новой, исследованиям в этой области посвящено множество научных и научно-практических работ. Тем не менее на текущий момент универсального и достаточно эффективного решения пока не предложено.

Простейшее улучшение ручной настройки алгоритмов – поиск по сетке. Дальнейшее улучшение – использование алгоритмов однокритериальной и многокритериальной оптимизации, где параметры метода машинного обучения являются объектными переменными, а в качестве целевой функции рассматривается эффективность применения метода машинного обучения. *Tužar* в своей работе использует дифференциальную эволюцию для многокритериальной оптимизации совместно с алгоритмом машинного обучения. *Kohavi* и *John* вели поиск подходящих параметров алгоритма C4.5 для построения деревьев решений. Согласно результатам, оптимизированные значения параметров алгоритма в большинстве случаев обеспечивают лучшую либо неуступающую точность решения задач (в частности, задач классификации и регрессии). Похожие эксперименты проводились Младеничем для поиска параметров при решении задачи пост-пруннинга дерева решений. Оптимизируемым критерием выступала точность классификации дерева решений, вычисленная по 10-кратной кросс-валидации. *Bohanec* и *Bratko* представили

алгоритм OPT, который на каждой итерации искал дерево решений, обеспечивающее наибольшую точность классификации среди всех деревьев того же размера. *Bergstra* использовал случайный поиск и алгоритм «Древовидная оценка Парзена» для поиска параметров нейронных сетей.

Работы некоторых авторов посвящены оптимизации параметров метода опорных векторов (*support vector machine, SVM*). *Rossi* и *Carvalho* провели сравнение четырех алгоритмов оптимизации параметров данного метода: генетический алгоритм, алгоритм клонируемой селекции, муравьиный алгоритм, алгоритм роя частиц. В некоторых случаях алгоритм *SVM* с параметрами по умолчанию оказался более эффективен, чем с оптимизированными параметрами. *Lessmann, Stahlbock* и *Crone* оптимизировали параметры алгоритма *SVM* с помощью генетического алгоритма. В сравнении с поиском по решетке генетический алгоритм обеспечил лучшие и более стабильные результаты. В работах *Almeida* и *Leung* использовались эволюционные алгоритмы для инициализации параметров нейронных сетей.

В последние годы активно развивается Красноярская научная школа. Наиболее известные и значимые результаты в области разработки эволюционных алгоритмов получены научной школой Е. С. Семенкина. В частности, Ш. А. К. Ахмедовой был разработан коллективный алгоритм оптимизации, комбинирующий в себе различные бионические алгоритмы. Р. Б. Сергиенко предложил коэволюционный алгоритм многокритериальной оптимизации. Разрабатываемые данной научной школой эволюционные алгоритмы многокритериальной оптимизации используются для оптимизации параметров алгоритмов машинного обучения, таких как нейронные сети (Ш. А. К. Ахмедова, К. Ю. Брестер), нечеткая логика (Р. Б. Сергиенко), генетическое программирование (Е. А. Сопов) и др. Эти алгоритмы используются для решения различных практических задач: распознавание эмоций человека по аудиозаписи и видеозаписи лица (М. Ю. Сидоров), выбор эффективных вариантов системы управления космическим аппаратом (М. Е. Семенкина) и мн. др.

Несмотря на то, что тема исследована большим количеством ученых и специалистов, исчерпывающего решения проблемы не предложено. Более того, появляются новые задачи, методы и модели машинного обучения, для которых также требуется разработка методов автоматизированного проектирования. Следовательно, разработка методов автоматизированной настройки алгоритмов машинного

обучения в целом и нейронных сетей в частности является актуальной научно-технической задачей.

Монография – результат научно-исследовательской работы авторов. Некоторые ее положения и частные результаты представлялись на различных всероссийских и международных научных конференциях и опубликованы в научных изданиях из списка ВАК РФ, включая индексируемые в базах *Scopus* и *Web of Science*. Программные реализации описанных алгоритмов имеют государственную регистрацию в Роспатенте, а сами подходы использовались при решении практических задач в рамках научных грантов, государственных заданий, проектов ФЦП и РФФИ и др. И. А. Ивановым в 2017 году по данной теме защищена диссертация на соискание ученой степени кандидата технических наук.

В данной монографии представлен анализ основных подходов к решению задачи распознавания эмоций, алгоритмов оптимизации и машинного обучения, включая методы глубинного обучения, а также предложен оригинальный коэволюционный алгоритм многокритериальной оптимизации. Даны результаты исследования эффективности коэволюционного алгоритма и его применения для решения задачи проектирования ансамбля классификаторов и отбора информативных признаков в задачах машинного обучения. Для распознавания изображений в рамках решения задачи человеко-машинного взаимодействия представлены конволюционная нейронная сеть с гибридным алгоритмом обучения на основе эволюционного алгоритма оптимизации и обобщенный метод решения задач классификации, включающих использование гетерогенных аудио-, видеоданных.

Глава 1

СИСТЕМНЫЙ АНАЛИЗ ПРОБЛЕМЫ ПРИМЕНЕНИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ И ОПТИМИЗАЦИИ В ЗАДАЧАХ ЧЕЛОВЕКО-МАШИННОГО ВЗАИМОДЕЙСТВИЯ

В первой главе рассмотрены известные методы и модели машинного и глубинного обучения, представлена задача человеко-машинного взаимодействия и обзор подходов к ее решению.

1.1. ОБЗОР СОВРЕМЕННЫХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ, КЛАССИФИКАЦИИ И ОПТИМИЗАЦИИ

Машинное обучение – обширный раздел искусственного интеллекта, посвященный разработке алгоритмов для обучения машин (алгоритмов, программных систем или аппаратно-программных комплексов) решению практических задач [4, 26]. Машинное обучение находится на стыке дисциплин, таких как математическая статистика, методы оптимизации, информатика. Кроме того, практическая направленность машинного обучения связывает его со многими другими областями человеческих знаний, на первый взгляд никак не связанными с математикой и вычислениями. К примеру, медицинская информационная система, способная автоматически ставить диагноз пациента по входным симптомам, относится к приложениям машинного обучения, но для создания такой системы, наряду со знаниями в области математических алгоритмов, требуются также знания в предметной области решаемой задачи – медицине. На сегодняшний день сфера применения алгоритмов машинного обучения стала столь широка, что данная дисциплина стала связана с большим количеством технических и гуманитарных отраслей человеческой деятельности.

Различают два типа машинного обучения – *индуктивное* и *дедуктивное*. Задача индуктивного обучения состоит в выявлении общих закономерностей в данных с целью их систематичного описания либо с целью прогнозирования будущих данных. При дедуктивном обучении создается некоторая общая модель на основании знаний экспертов предметной области, которая используется для вывода конечных заключений. Приложения дедуктивного обучения относят к отдельной области экспертных систем [14, 27], поэтому на практике под машинным обучением обычно понимают индуктивное обучение.

Индуктивное обучение, которое далее мы будем называть просто машинным обучением, как следует из его определения, неразрывно связано с данными. Данные, как правило, представляют собой некоторые прецеденты из предметной области, организованные в виде таблиц, в которых каждая строка представляет собой вектор, описывающий отдельный прецедент. Например, в задаче медицинской диагностики данные представляют собой таблицу симптомов пациентов, в которой каждая строка соответствует отдельному пациенту, а каждый столбец – отдельному симптому. Таким образом, машинное обучение тесно связано с другой развивающейся дисциплиной – интеллектуальным анализом данных (*data mining*) [7, 12], главной целью которого является обнаружение в данных ранее неизвестных и нетривиальных закономерностей.

Машинное обучение по типу использования информации подразделяется на несколько подклассов:

1. Обучение с учителем (*supervised learning*).
2. Обучение без учителя (*unsupervised learning*).
3. Частичное обучение (*semi-supervised learning*).
4. Обучение с подкреплением (*reinforcement learning*).
5. Динамическое обучение (*online learning*).
6. Активное обучение (*active learning*).

Обучение с учителем является наиболее распространенным вариантом машинного обучения в современных практических приложениях. Данный тип обучения работает с данными, организованными в виде структуры «объект – метка». Задача состоит в обучении алгоритма, восстанавливающего некую зависимость между признаками объекта и метками. При этом исходные данные разбиваются на непересекающиеся выборки – обучающую и тестовую. Различают несколько типов задач обучения с учителем:

Задача классификации (*classification*) [15]. В данной задаче конечное множество возможных меток объектов, называемых метками

классов, или просто классами. Задача алгоритма обучения состоит в правильном отнесении объекта к одному из классов. Качество работы алгоритма определяется ошибкой классификации, т. е. долей объектов тестовой выборки, отнесенных к неверному классу.

Задача регрессии (regression) [31] отличается от задачи классификации тем, что меткой каждого объекта служит действительное число, следовательно, множество возможных меток неограниченно. Алгоритм обучения аппроксимирует некоторую функциональную зависимость числовой метки объекта от его признаков.

Задача прогнозирования (forecasting) [132]. Объектами являются значения некоторого параметра (вектора параметров), расположенные по оси времени. Совокупность таких объектов называют временным рядом, а саму задачу – прогнозированием временных рядов. Задача алгоритма обучения – на основании имеющихся объектов сделать прогноз на будущее.

Обучение без учителя использует данные, в которых не заданы метки объектов, т. е. каждый объект представляет собой вектор значений признаков либо вектор расстояний в признаковом пространстве до остальных объектов выборки. Цель алгоритмов обучения без учителя – поиск зависимостей между объектами на основании данных об их признаках. Различают следующие задачи обучения без учителя:

Задача кластеризации (clustering) [16] заключается в разбиении выборки объектов на группы таким образом, чтобы объекты внутри группы были схожи по некоторым признакам, а объекты разных групп отличались по этим признакам. Так как в данной задаче нет меток классов, как в задаче классификации, критерий качества кластеризации может быть задан как отношение среднего межкластерного и среднего внутрикластерного расстояния между объектами. Чем больше среднее расстояние между кластерами и чем меньше среднее расстояние между объектами одного кластера, тем лучше алгоритм кластеризации разделил объекты.

Задача фильтрации выбросов (outlier detection) [18] состоит в поиске нетипичных объектов выборки, отличных от других. Данная задача может быть как конечной, так и вспомогательной при решении задач обучения с учителем. Например, в практической задаче обнаружения бракованных деталей на предприятии поиск объектов выборки с нетипичными признаками является самоцелью, тогда как при решении задачи классификации на основе неточных,

ошибочных данных поиск и исключение из выборки выбросов могут послужить повышению точности конечной системы обучения с учителем.

Задача сокращения размерности (dimensionality reduction) [2] заключается в применении некоторых преобразований над данными, переводящих исходные признаки к меньшему числу новых признаков без потери информации об объектах выборки. Алгоритмы, решающие данную задачу, объединены под названием алгоритмов факторного анализа. Задача сокращения размерности может также быть отнесена к обучению с учителем, так как есть подкласс задач сокращения размерности, называемый отбором признаков, в котором исходные признаки не трансформируются в новые, а лишь выбираются наиболее информативные из числа имеющихся.

Задача заполнения пропусков в данных (missing values imputation) [119] актуальна при работе с выборками, в которых присутствуют пропущенные значения. Цель данной задачи состоит в прогнозировании пропущенных значений по имеющимся данным.

Частичное обучение (semi-supervised learning) имеет дело с частично размеченными данными, т. е. метки даны лишь для некоторых объектов выборки [57]. Например, в базе данных клиентов указан возраст одной трети людей, необходимо предсказать возраст остальной части людей, для которых он не указан.

Обучение с подкреплением (reinforcement learning) отличается тем, что в нем роль объектов играют пары «состояние – действие» (*state – action*), а в качестве метки выступает реакция окружающей среды на произведенное действие, характеризующая правильность действия [136]. Обучение с подкреплением является более сложной версией обучения с учителем, так как реакция среды может быть не мгновенной, а достаточно отдаленной во времени, что усложняет прогнозирование. Обучение с подкреплением нашло обширное применение в робототехнике, где робот (агент) в начале своего обучения ничего не знает об окружающей среде, но постепенно учится эффективно взаимодействовать с ней, получая и анализируя реакцию среды на свои действия.

В динамическом обучении (online learning) объекты поступают один за другим, а не все сразу. В связи с этим алгоритму приходится обрабатывать каждый объект по отдельности и дообучаться с учетом новых знаний о новом объекте. Модели динамического обучения крайне важны с точки зрения практики, так как в реальных задачах,

как правило, данные не представлены в полном объеме сразу, а поступают порциями с течением времени.

Наконец, при *активном обучении (active learning)* объекты поступают на вход алгоритма не случайно, а в определенной последовательности, заложенной в алгоритм, которая позволяет ему эффективнее обучаться. Этот тип обучения связан с областью планирования эксперимента [1].

В данной монографии акцент делается на исследование алгоритмов обучения с учителем и на решение задач классификации, эти подходы будут рассмотрены более детально.

Формальная математическая постановка задачи классификации выглядит следующим образом [2]. Пусть X – множество объектов, Y – конечное множество меток классов этих объектов. Пусть существует некоторая неизвестная зависимость $f: X \rightarrow Y$, связывающая объекты и их метки. Данная зависимость известна лишь для объектов конечной обучающей выборки $(X_{\text{обуч}} \in X) = \{(\bar{x}_1, y_1, \bar{x}_2, y_1, \dots, \bar{x}_n, y_n)\}$. Задача – построить алгоритм, называемый также решающим правилом, способный классифицировать произвольный объект $\bar{x} \in X$, $\bar{x} \notin X_{\text{обуч}}$.

Алгоритмы классификации неразрывно связаны с данными, на которых они строятся и обучаются. Данные чаще всего представлены в виде признакового описания объектов классификации. При этом признаки могут быть в количественной, порядковой либо номинальной шкале. Другой вариант представления данных, использующийся значительно реже, – с помощью матрицы расстояний между объектами. Всегда существует возможность перехода от одного типа представления данных к другому путем использования какой-либо метрики, т. е. способа вычисления расстояния между объектами на основе их признаков.

Ключевым понятием, неразрывно связанным с классификацией, является понятие признакового пространства. Если задан набор признаков f_1, f_2, \dots, f_m , которыми описывается объект, то множество $X = D_{f_1} \times D_{f_2} \times \dots \times D_{f_m}$ называют признаковым пространством, а D_{f_i} – областью определения (множество допустимых значений) признака f_i .

Задачи классификации различают по типам и количеству классов:

1. Двухклассовая классификация.
2. Мультиклассовая классификация [137] – в задаче больше двух классов.

3. Пересекающиеся/непересекающиеся классы – каждый объект может относиться только к одному классу или же к нескольким классам.

4. Нечеткая классификация [90, 143] – требуется определить степень принадлежности объекта каждому из классов.

Существует несколько общепринятых способов оценки качества алгоритмов классификации:

1. Разделение выборки на обучающую и тестовую – модель строится по данным из обучающей выборки и тестируется по тестовой выборке. Критерием качества выступает ошибка классификации, либо среднеквадратичная ошибка в случае задачи регрессии, либо иной критерий.

2. Кросс-проверка (*cross-validation*) – исходные данные случайно разделяются на m подвыборок, процедура обучения и тестирования модели проводится m раз: в первый раз модель обучается на всех подвыборках, кроме первой, тестируется на первой; затем обучается на всех подвыборках, кроме второй, тестируется на второй и т. д. Такой процесс называется m -кратной кросс-проверкой.

3. Скользящий экзамен (*leave-one-out experiment*) – является n -кратной кросс-проверкой, где n – объем выборки. То есть сперва модель строится на всех объектах выборки, кроме первого, проверяется на первом; затем данная процедура повторяется для всех объектов выборки.

Существующие алгоритмы классификации могут быть объединены в несколько групп по принципу их действия:

1. Классификаторы на основе сходства объектов.

2. Алгоритмы статистической классификации.

3. Классификаторы на основе делимости классов в признаковом пространстве.

4. Алгоритмы логической классификации.

5. Нейронные сети.

Алгоритмы на основе сходства объектов основаны на гипотезе компактности, в которой предполагается, что объекты одного класса чаще всего похожи друг на друга, а объекты разных классов отличны. Схожесть объектов в n -мерном признаковом пространстве может вычисляться по различным метрикам:

Евклидово расстояние

$$d_E(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}. \quad (1.1)$$

Манхэттенское расстояние

$$d_{Man}(x, y) = \sum_{i=1}^n |x_i - y_i|. \quad (1.2)$$

Чебышевское расстояние

$$d_{Ch}(x, y) = \max_i (|x_i - y_i|). \quad (1.3)$$

Расстояние Минковского p -го порядка

$$d_{Mink}(x, y) = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p}. \quad (1.4)$$

Наиболее известные алгоритмы этой группы следующие:

Метод построения эталонов. По обучающей выборке вычисляются точки-эталоны, являющиеся центрами объектов каждого класса в признаковом пространстве, по формуле:

$$E_i^k = \frac{1}{n^k} \sum (x_i : f(\bar{x}) = k); i = \overline{1, m}; k = \overline{1, K}, \quad (1.5)$$

где E^k – эталон k -го класса; n^k – число объектов выборки k -го класса; m – размерность признакового пространства; K – число классов.

По определенной метрике рассчитывается расстояние от классифицируемого объекта до каждого из эталонов, объект относят к тому классу, расстояние до эталона которого минимально.

Метод k ближайших соседей [38]. Для этого алгоритма данные должны быть представлены в виде матрицы расстояний между объектами выборки, вычисленных по определенной метрике. Рассматриваются классы объектов, ближайших к классифицируемому. Объект относится к тому классу, который наиболее часто встречается среди его соседей.

При решении практических задач данным методом очень важно выбрать правильную метрику вычисления расстояния между объектами, а также значение параметра k . При слишком маленьком значении параметра (например, $k = 1$) алгоритм становится подвержен негативному влиянию выбросов, при слишком высоком значении k в расчеты включается слишком много соседних объектов, что также может негативно сказаться на качестве классификации, так как среди

соседей классифицируемого объекта может оказаться много объектов другого класса.

Метод потенциальных функций [3]. Данный метод построен на физическом принципе потенциала электрического поля заряженной частицы. Рассчитывается расстояние от классифицируемого объекта до каждого объекта обучающей выборки. Решающее правило строится, как в методе ближайших соседей, разница состоит в том, что объект выборки обладает некоторой меткой важности («зарядом») относительно классифицируемого объекта.

Алгоритмы статистической классификации основаны на оценивании плотности распределения классов по выборке. В зависимости от способа оценивания различают алгоритмы с параметрическим и непараметрическим оцениванием плотности, а также оцениванием плотности как смеси параметрических распределений. Известные алгоритмы этой группы:

Наивный байесовский классификатор [116]. Данный метод основан на предположении, что признаки, которыми описываются объекты выборки, статистически независимы. Данное предположение существенно облегчает задачу оценивания плотности распределения, так как вместо n -мерной плотности необходимо оценить n одномерных плотностей. Плотности могут оцениваться как параметрическим, так и непараметрическим способом. По правилу Байеса находятся апостериорные вероятности каждого из K классов при условии измерения признака x классифицируемого объекта:

$$P(i|x) = \frac{f(x|i) * P(i)}{\sum_{j=1}^m f(x|i) * P(j)}, i = \overline{1, K}, \quad (1.6)$$

где $f(x|i)$ – оценка условной плотности распределения признака x для i -го класса; $P(i)$ – оценка априорной вероятности класса. Решающее правило для классифицируемого объекта x выглядит следующим образом:

$$i^* = \arg \max_i P(i|x), i = \overline{1, K}. \quad (1.7)$$

Метод парзеновского окна использует непараметрическое оценивание плотности [17] распределения классов по имеющейся выборке, следовательно, в нем не выдвигаются гипотезы о структуре

функции плотности распределения. Решающее правило классификации объекта x выглядит следующим образом:

$$i^* = \arg \max_i \left(\lambda_i \cdot \sum_{j=1}^n [y_j = i] \cdot K \left(\frac{d(x, x_j)}{h} \right) \right), i = \overline{1, K}, \quad (1.8)$$

где λ_j – цена правильного ответа для класса i ; n – объем выборки; y_j – класс j -го объекта; $K(t)$ – ядерная функция; $d(x, x_j)$ – расстояние между классифицируемым объектом x и объектом x_j ; h – ширина окна.

ЕМ-алгоритм (expectation-maximization) [65] оценивает плотность как смесь параметрических распределений. В данном алгоритме итеративно выполняются два этапа: этап оценки (*estimation*), на котором рассчитывается ожидаемое значение функции правдоподобия, и этап максимизации (*maximization*), на котором вычисляются параметры функции правдоподобия, доставляющие ей максимум.

Следующую группу составляют *классификаторы на основе разделимости классов*. Данные алгоритмы строят разделяющую поверхность в признаковом пространстве, разделяющую объекты на непересекающиеся классы. Наиболее известные методы данной группы следующие:

Линейный дискриминант Фишера [125], также известный как линейный дискриминантный анализ, применим, если выборка удовлетворяет следующим гипотезам: классы распределены по нормальному закону, и матрицы ковариаций классов равны. Линейный дискриминант Фишера является упрощением квадратичного дискриминанта. В случае двух классов в двумерном пространстве разделяющей поверхностью, построенной с помощью этого метода, будет прямая. В случае большего числа классов разделяющая поверхность будет кусочно-линейной.

Логистическая регрессия [85]. Для случая двух классов строится линейный алгоритм классификации с решающим правилом вида:

$$\log \text{Reg}(x, w) = \text{sign}(\sum_{j=1}^m w_j x_j - w_0) = \text{sign}(\langle x, w \rangle), \quad (1.9)$$

где w_j – вес j -го признака; w_0 – порог принятия решения, w – вектор весов; $\langle x, w \rangle$ – скалярное произведение вектора весов и признаков объекта. Задача обучения алгоритма логистической регрессии состоит в нахождении оптимального вектора весов w , минимизирующего функцию потерь вида:

$$L(w) = \sum_{i=1}^n \ln(1 + \exp(-y_i \langle x_i, w \rangle)). \quad (1.10)$$